JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 2 (2014), Issue 2 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Original Papers

Viewpoints

Original Paper

A Fuzzy-Match Search Engine for Physician Directories

Majid Rastegar-Mojarad[1*], MS; Christopher Kadolph[1*], BS; Zhan Ye[1], PhD; Daniel Wall[1*], BS; Narayana Murali[2], MD; Simon Lin[3], MD

[1]Marshfield Clinic Research Foundation, Biomedical Informatics Research Center, Marshfield, WI, United States

[2]Marshfield Clinic, Department of Nephrology, Marshfield, WI, United States

[3]The Research Institute at Nationwide Children's Hospital, Columbus, OH, United States

[*]these authors contributed equally

Corresponding Author:
Simon Lin, MD
The Research Institute at Nationwide Children's Hospital
575 Children's Crossroad
Columbus, OH, 43017
United States
Phone: 1 614 355 6629
Fax: 1 614 355 5601
Email: Simon.Lin@NationwideChildrens.org

Abstract

Background: A search engine to find physicians' information is a basic but crucial function of a health care provider's website. Inefficient search engines, which return no results or incorrect results, can lead to patient frustration and potential customer loss. A search engine that can handle misspellings and spelling variations of names is needed, as the United States (US) has culturally, racially, and ethnically diverse names.

Objective: The Marshfield Clinic website provides a search engine for users to search for physicians' names. The current search engine provides an auto-completion function, but it requires an exact match. We observed that 26% of all searches yielded no results. The goal was to design a fuzzy-match algorithm to aid users in finding physicians easier and faster.

Methods: Instead of an exact match search, we used a fuzzy algorithm to find similar matches for searched terms. In the algorithm, we solved three types of search engine failures: "Typographic", "Phonetic spelling variation", and "Nickname". To solve these mismatches, we used a customized Levenshtein distance calculation that incorporated Soundex coding and a lookup table of nicknames derived from US census data.

Results: Using the "Challenge Data Set of Marshfield Physician Names," we evaluated the accuracy of fuzzy-match engine–top ten (90%) and compared it with exact match (0%), Soundex (24%), Levenshtein distance (59%), and fuzzy-match engine–top one (71%).

Conclusions: We designed, created a reference implementation, and evaluated a fuzzy-match search engine for physician directories. The open-source code is available at the codeplex website and a reference implementation is available for demonstration at the datamarsh website.

(JMIR Med Inform 2014;2(2):e30) doi:10.2196/medinform.3463

KEYWORDS

Fuzzy-Match; Levenshtein Distance; Physician Name; Physician Directory

Introduction

A primary functionality of the website of a physician group practice is a search engine where patients can enter a physician's name and find more information about the physician's practice, credentials, and appointment phone number. Name-based searching seems to be a simple task, but various types of spelling mismatches caused by typographical errors, phonetic spelling variations, and nicknames can make the task difficult. Failure to find a physician on the provider's website can create a frustrating experience for the patient and potential loss of business for the provider.

We surveyed the websites of the ten largest medical groups [1], and found none of them allowed mismatched characters in the

entered name. 7 of the 10 search engines allowed autocomplete, which tries to finish the rest of the characters based on what has already been typed. However, current implementations of auto-complete require 100% match in the already typed fragments, any mismatches will end up with no results. The Google search engine does allow fuzzy-match, but it is not specific to the physician directory on a provider's website. Consequently, a general Google search of a physician's name might lead to websites other than the provider's. As such, the Google search does not provide an integrated patient experience at the provider's website. Currently, there are no open-source solutions of a fuzzy-match search engine for physician directories.

To improve upon current and severely limited provider search engines, we conducted a heuristic analysis of the search log. A common mismatch can be caused by typographical errors. For example, "Smith" is entered as "Smitj", because the "j" key is adjacent to the "h" key. As more people are searching websites using smaller touch-screen devices such as smartphones, typographical errors resulting from adjacent keys are becoming more common. Levenshtein distance based methods, as previously used in matching drug names and chemical names [2,3], can be effective in correcting this type of error. Levenshtein distance is a measure of the similarity between two strings. The distance is the number of deletions, insertions, or substitutions required to transform one string to the other. For instance, the Levenshtein distance between "Smith" and "Smitj" is one, whereas an exact match results in a distance of zero.

Another type of mismatch is caused by phonetic variations in names. For instance, "Smith" and "Smyth" are pronounced the same but spelled differently. Sound-based encoding methods such as Soundex and Metaphone were designed to solve the phonetic variation in names. In 1918, Robert Russell developed the first Soundex system and subsequently, several implementations were devised. Soundex encodes [4] names based on their sound, so that names with close pronunciation get the same code. For example, both "Smith" and "Smyth" are coded as "S530". One problem with Soundex is that it returns many approximate matches, with most being far from the searched-for name [5]. Beidar and Morse [5] developed the Beider-Morse Phonetic Matching system for decreasing the number of approximate matches by removing irrelevant ones. Lawrence Phillips upgraded the Soundex system in 1990 and developed Metaphone [6], which produces more accurate encoding of names that sound similar. Further development of Double Metaphone [6] enabled two codes for a single name to account for different kinds of spelling variations. Double Metaphone also improved the match of non-English names. However, implementations of Soundex or Metaphone are usually outside of the aforementioned Levenshtein distance framework.

A third type of variation is caused by nicknames. For instance, "Bill" might exist in the directory as "William." Since "Bill" and "William" do not sound, nor are spelled alike, nicknames pose another challenge for name searches. Nicknames cannot be resolved by distance-based match or sound-based match. None of the search engines at the ten largest medical groups had a good solution for nicknames. We proposed to use a nickname lookup table [7] derived from the United States (US) census data to solve this problem, where we also incorporated it in the Levenshtein distance framework.

In the medical informatics literature, the approximate match of patient names has been studied extensively. Both phonetic name matching and Levenshtein distance based methods were reported [8,9]. Peter Christen [10] presented a comprehensive review on the name matching algorithms; however, there have been no reports of an integrated solution that simultaneously addresses all three kinds of mismatches.

Marshfield Clinic has more than 800 providers with diverse first and last names. A fast and effective "Find a doctor" engine is critical to the business operation. From the log file of the "Find a doctor" webpage at Marshfield Clinic, we observed that 26% of the 9072 searches in July 2013 yielded no results. To aid patients in finding the wanted provider easier and faster, we suggest a list of providers' name that are similar to the search term. As a patient enters the name of the desired physician, our system provides a list of suggestions that helps the user, even if they do not know the correct spelling of the wanted physician's name. Unlike most available systems, our system applies approximate search instead of exact match search for finding similar names. This article presents an open-source solution, demonstrates the implementation, and evaluates the effectiveness of a fuzzy search engine for physician directories. The novelty in our system is that it is the first open-source search engine for physician directories that solves all three kinds of spelling mismatches: typographical errors, phonetic variations, and nicknames.

Methods

In our application, it was imperative to find the closest physician's name in the directory to the entered search term. First, we performed some preprocessing steps. We removed common prefixes and suffixes in the string, such as Dr, MD, FACS, etc. Then, to solve all three kinds of mismatches in a unified framework, we customized the Levenshtein distance method. Refer to Textbox 1. for the assigned cost for each operation.

Textbox 1. Cost of operation.

1. Cost of deletion is:

I. 4 if the letter is ‘a’, ‘e’, ‘i’, ‘o’, or ‘u’

II. 4 if the letter is the same as the previous letter (repetitive letters)

III. otherwise 5

2. Cost of substitution is:

I. 3 if both letters have a similar sound. Here, we used Soundex to determine whether two letters have the same sound. For example, we assumed that ‘d’ and ‘t’ have the same sound, because they have the same code in Soundex.

II. 3 if they are adjacent on keyboard. We took eight surrounding keys for each character and assigned them with lower penalties to accommodate typographical errors.

III. otherwise 4

Additionally, we used the nickname lookup table to expand the match to the physician directory. Each nickname is assigned with a matching likelihood. For instance, “William” has a 0.9 chance of being called “Bill” and 0.45 chance of being called “Will”. We also incorporated the probability in the final matching score.

To evaluate the performance of the method, we chose 100 recently searched terms from the Marshfield Clinic website’s current search engine (uses exact match approach) log file, which did not return any results. Using human intelligence, we identified the correct physicians name in the Marshfield Clinic directory for 68 of the searched terms. We call this gold-standard data set the “Challenge Data Set of Marshfield Physician Names”. Ten examples in this data set are shown in Table 1.

Table 1. Example data in the “Challenge Data Set of Marshfield Physician Names”.

Search term entered by patient	Actual name in the directory
alvarex	Maria Alvarez
carrie tull	Carie Tull
Ceasar gonzaga	Caesar Gonzaga
phillip zickerman	Philip Zickerman
reinhardt	Richard Reinhart
roedrick koehler	Roderick Koehler
rousch	Stephen Roush
scott erickwon	Scott Erickson
STEVEN TOOTHACKER	Stephen Toothaker
tim swan	Timothy Swan

To compare diversity of the names of US physicians versus general US population, a list of 1,048,576 physician names was obtained from the National Provider Identifier Registry of 2013 [11]. The names of the general US population were obtained from the website of the US Census Bureau [12]. Because the 1990 census is the latest one containing statistics with both first and last names, we used it in this study.

Results

It is important to note that physician names are more diverse than those of the general US population. By comparing the nationwide physician names listed in the National Provider Identifier registry with the general US population, we confirmed that the physician names are less common than names in the general US population (Figure 1). For instance, to cover 70% of the last names, 9028 names need to be included for the general US population, whereas 40,014 names need to be included for physician names. The same is true for first names, but to a lesser extent (Figure 1). Consequently, less common names can be more challenging to spell correctly. To assess the statistical significance, we utilized two sample Kolmogorov-Smirnov (K-S) tests on the two cumulative distributions from each of the three graphs in Figure 1. The results show P values <.001, which indicates there are significant differences between the two distributions of the cumulative coverage of physician last names, male first names, and female first names, respectively.

Less common names, combined with phonetic variations, nicknames, and typographical errors, pose challenges to search engines at a group practice provider’s website. We researched the “Find a Doctor” webpages at the top 10 medical groups in the United States (Table 2). None of the websites allowed fuzzy-match of physicians’ names. While 7 out of 10 websites have the autocompletion feature, none allow any mismatches in the name search query.

Table 2. "Find a doctor" search engines at top ten medical groups in the United States [3].

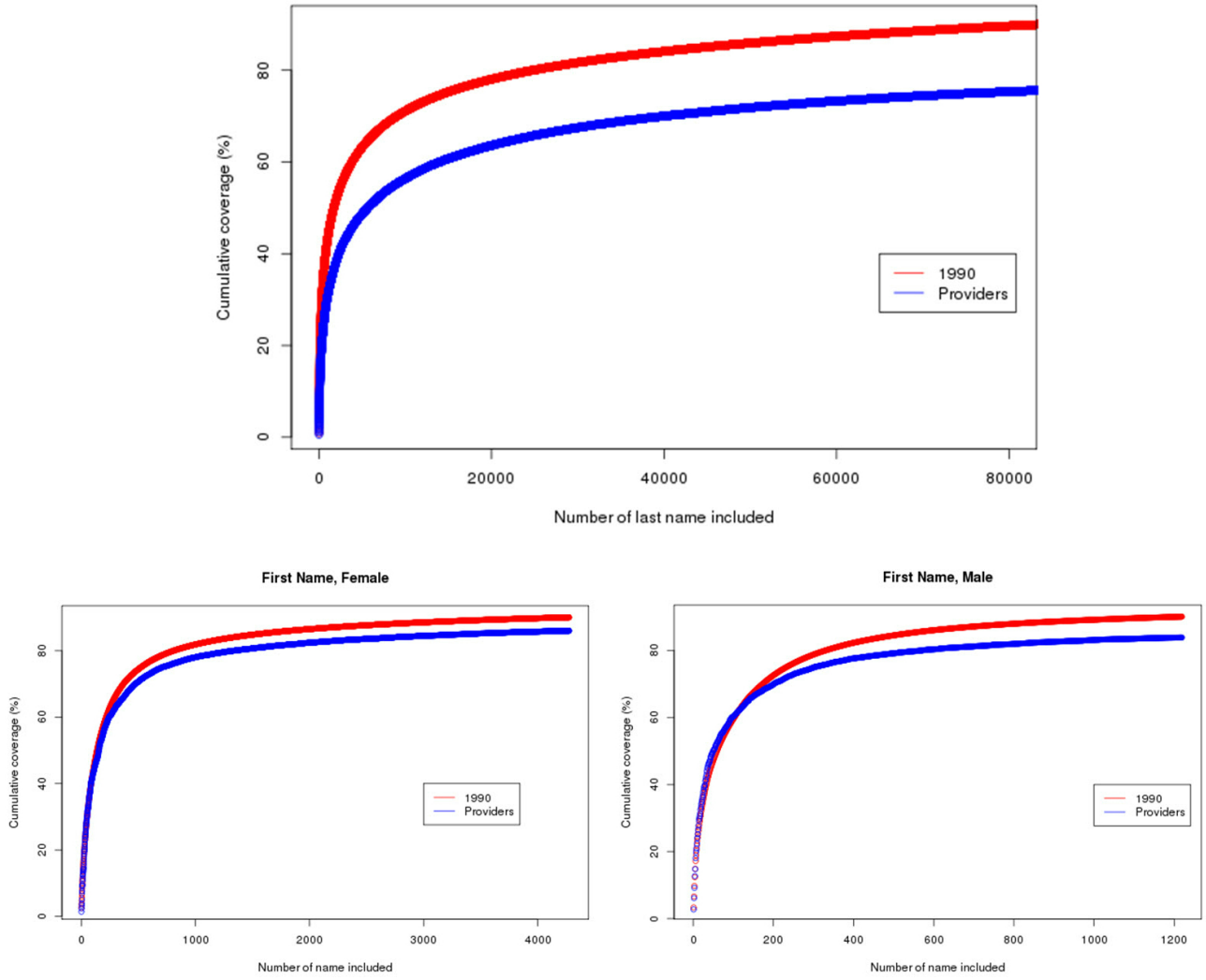
Medical group	Headquarters	Offices	Physicians	Auto completion
Kaiser Permanente Medical Group	Santa Clara, CA	484	7842	No
Cleveland Clinic	Cleveland, OH	173	1472	Yes
Henry Ford Medical Group	Detroit, MI	218	1224	Yes
IU Health Physicians	Indianapolis, IN	267	1202	No
University Washington Physicians	Seattle, WA	181	1199	Yes
Mercy Springfield	Springfield, MO	349	1115	Yes
North Shore Long Island Jewish Syosset	Syosset, NY	259	1044	Yes
Carolinas Primary Care	Loris, SC	236	1024	Yes
Aurora Medical Group	Sheboygan, WI	206	1013	Yes
Novant Medical Group	Winston-Salem, NC	245	923	No

The "Challenge Data Set of Marshfield Physician Names" was used to evaluate the performance of a fuzzy match search engine. In the first comparison, the accuracy of fuzzy-match with Soundex algorithm was compared. Table 3 illustrates the results of this experiment. It should be emphasized that the current search engine returned "no result" for these 68 search terms. In the second evaluation, a comparison was done using simple Levenshtein distance in fuzzy-match versus customized Levenshtein distance. For similarity-based search methods, more than one result could be returned. As such, we also compared the efficiency of returning top ten matches versus top one match. The results suggest the top-one match already significantly outperforms Soundex, and the top-ten matches can further improve the retrieval performance.

Table 3. Comparing accuracy of Soundex, Levenshtein Distance (LD), and Fuzzy-Match on the Challenge Data Set of Marshfield Physician Names (N=68).

Search Engines	# Found	Percentage
Default Search Engine	0	0%
Soundex	16	24%
Fuzzy-match with simple LD (top one)	40	59%
Fuzzy-match with customized LD (top one)	48	71%
Fuzzy-match with simple LD (top ten matches)	52	77%
Fuzzy-match with customized LD (top ten matches)	61	90%

Figure 1. Physician's first name and last name, comparing with general US population.



Discussion

Principal Findings

This study focuses on the search engine used by patients to search the physician directory at a provider's website. The same methods can be used to search any name directory system; for example, a directory of professors and staff members in the school of art and science of a university. It can also be used for Intranet searches. Staff members at Marshfield Clinic relate anecdotes about the inability to find the pager number for a physician in the Intranet directory, because they could not get the first character of the name spelled correctly. For example, "Przybylinski" (pronounced as "Shibilinski") cannot be found under the directory using the starting letter "S"; however, using the fuzzy search engine presented in this paper, a top match can be found. The "Challenge Data Set of Marshfield Physician Names", although small, can also be used in the future as a benchmark data set to test search engines of physician names.

Conclusions

We designed and evaluated a fuzzy-match search engine for physician directories. The open-source code is available at Codeplex web site [13] and a reference implementation is demonstrated at datamarsh website under FuzzyMatch [14].

Acknowledgments

The authors would like to thank Madalyn Minervini and DeeAnn Polacek for their manual creation of the of the "Challenge Data Set of Marshfield Physician Names" from the search logs; William Hogg for extracting data from the search log; Robert Moritz and John Tracy for helpful discussions of the search engine for physician directories; and Joe Finamore, Andrea Mahnke, and Po-Huang Chyou for project discussions. We thank Dr Ingrid Glurich for critical review and Marie Fleisner for editing. Majid Rastegar-Mojarad was funded through philanthropic support of Marshfield Clinic Research Foundation's "Dr John Melski Endowed Physician Scientist" Award to Dr Simon Lin.

Conflicts of Interest

None declared.

References

1. SK&A, a Cegedim Company. 2013. SK&A's 50 Largest Medical Groups URL: http://www.skainfo.com/health_care_market_reports/largest_medical_groups.pdf [accessed 2014-04-08] [WebCite Cache ID 6OghjdsYx]
2. Wang JF, Li ZR, Cai CZ, Chen YZ. Assessment of approximate string matching in a biomedical text retrieval problem. Comput Biol Med 2005 Oct;35(8):717-724. [doi: 10.1016/j.compbiomed.2004.06.002] [Medline: 16124992]
3. Peters L, Kapusnik-Uner JE, Nguyen T, Bodenreider O. An approximate matching method for clinical drug names. AMIA Annu Symp Proc 2011;2011:1117-1126 [FREE Full text] [Medline: 22195172]
4. National Archives. The Soundex Indexing System URL: http://www.archives.gov/research/census/soundex.html [accessed 2014-04-08] [WebCite Cache ID 6OghnAOIk]
5. Beider A. Avotaynu: International Review of Jewish Genealogy. Beider-Morse phonetic matching: an alternative to Soundex with fewer false hits URL: http://stevemorse.org/phonetics/bmpm.htm [accessed 2014-04-08] [WebCite Cache ID 6Oghd8uDP]
6. Philips L. Hanging on the Metaphone. Computer Language Magazine 1990;7(12).
7. Deron M. Most common nicknames for first names URL: http://deron.meranda.us/data/nicknames.txt [accessed 2014-04-08] [WebCite Cache ID 6OghrGbrq]
8. Grannis SJ, Overhage JM, McDonald C. Real world performance of approximate string comparators for use in patient matching. Stud Health Technol Inform 2004;107(Pt 1):43-47. [Medline: 15360771]
9. Levin HI, Levin JE, Docimo SG. "I meant that med for Baylee not Bailey!": a mixed method study to identify incidence and risk factors for CPOE patient misidentification. AMIA Annu Symp Proc 2012;2012:1294-1301 [FREE Full text] [Medline: 23304408]
10. Christen P. ICDM Workshops. 2006. A comparison of personal name matching: techniques and practical issues URL: https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=4063580 [accessed 2014-04-08] [WebCite Cache ID 6Oghv6UVU]
11. Office of the Secretary, HHS. Administrative simplification: adoption of a standard for a unique health plan identifier; addition to the National Provider Identifier requirements; and a change to the compliance date for the International Classification of Diseases, 10th Edition (ICD-10-CM and ICD-10-PCS) medical data code sets. Final rule. Fed Regist 2012 Sep 5;77(172):54663-54720. [Medline: 22950146]
12. United States Census Bureau. Does the U.S. Census Bureau provide any data by first names and surnames?. Frequently Asked Questions URL: https://ask.census.gov/faq.php?id=5000&faqId=3 [accessed 2014-04-08] [WebCite Cache ID 6Ogi0aohi]
13. A fuzzy match search engine for physician directories. 2014 Feb 13. URL: http://fuzzymatch.codeplex.com/ [WebCite Cache ID 6Thr7duze]
14. FMOQT: Fuzzy Match Online Query Tool. 2014. URL: http://datamarsh.org/FuzzyMatch [WebCite Cache ID 6ThrOGoBX]

Edited by G Eysenbach; submitted 09.04.14; peer-reviewed by M Rethlefsen, H Zhai; comments to author 20.08.14; revised version received 06.09.14; accepted 16.09.14; published 04.11.14.

Please cite as:
Rastegar-Mojarad M, Kadolph C, Ye Z, Wall D, Murali N, Lin S
A Fuzzy-Match Search Engine for Physician Directories
JMIR Med Inform 2014;2(2):e30
URL: http://medinform.jmir.org/2014/2/e30/
doi:10.2196/medinform.3463
PMID:25601050

Original Paper

Clinical Data Miner: An Electronic Case Report Form System With Integrated Data Preprocessing and Machine-Learning Libraries Supporting Clinical Diagnostic Model Research

Arnaud JF Installé[1,2], MSc, PhD; Thierry Van den Bosch[3], MD, PhD; Bart De Moor[1,2], MSc, PhD; Dirk Timmerman[3,4], MD, PhD

[1]Department of Electrical Engineering ESAT, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium

[2]iMinds Medical IT, Leuven, Belgium

[3]Department of Obstetrics and Gynecology, UZ Leuven, KU Leuven, Leuven, Belgium

[4]Department of Development and Regeneration, UZ Leuven, KU Leuven, Leuven, Belgium

Corresponding Author:
Arnaud JF Installé, MSc, PhD
Department of Electrical Engineering ESAT
STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics
KU Leuven
Kasteelpark Arenberg 10 - box 2446
Leuven, B-3001
Belgium
Phone: 32 16 328646
Fax: 32 16 321970
Email: arnaud.installe@esat.kuleuven.be

Abstract

Background: Using machine-learning techniques, clinical diagnostic model research extracts diagnostic models from patient data. Traditionally, patient data are often collected using electronic Case Report Form (eCRF) systems, while mathematical software is used for analyzing these data using machine-learning techniques. Due to the lack of integration between eCRF systems and mathematical software, extracting diagnostic models is a complex, error-prone process. Moreover, due to the complexity of this process, it is usually only performed once, after a predetermined number of data points have been collected, without insight into the predictive performance of the resulting models.

Objective: The objective of the study of Clinical Data Miner (CDM) software framework is to offer an eCRF system with integrated data preprocessing and machine-learning libraries, improving efficiency of the clinical diagnostic model research workflow, and to enable optimization of patient inclusion numbers through study performance monitoring.

Methods: The CDM software framework was developed using a test-driven development (TDD) approach, to ensure high software quality. Architecturally, CDM's design is split over a number of modules, to ensure future extendability.

Results: The TDD approach has enabled us to deliver high software quality. CDM's eCRF Web interface is in active use by the studies of the International Endometrial Tumor Analysis consortium, with over 4000 enrolled patients, and more studies planned. Additionally, a derived user interface has been used in six separate interrater agreement studies. CDM's integrated data preprocessing and machine-learning libraries simplify some otherwise manual and error-prone steps in the clinical diagnostic model research workflow. Furthermore, CDM's libraries provide study coordinators with a method to monitor a study's predictive performance as patient inclusions increase.

Conclusions: To our knowledge, CDM is the only eCRF system integrating data preprocessing and machine-learning libraries. This integration improves the efficiency of the clinical diagnostic model research workflow. Moreover, by simplifying the generation of learning curves, CDM enables study coordinators to assess more accurately when data collection can be terminated, resulting in better models or lower patient recruitment costs.

(JMIR Med Inform 2014;2(2):e28) doi:10.2196/medinform.3251

KEYWORDS

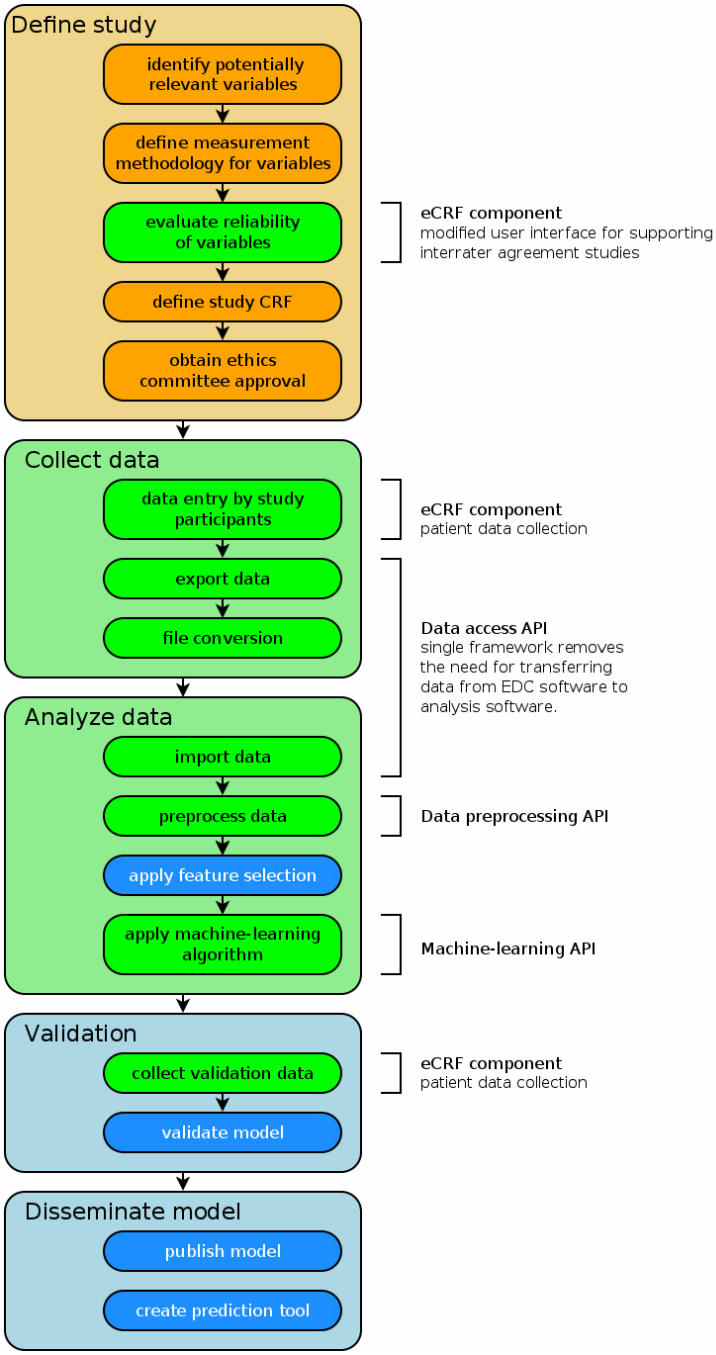
data collection; machine-learning; clinical decision support systems; data analysis

Introduction

Saving Lives With Early Detection

Many diseases, including cancer, may be cured or managed, if diagnosed sufficiently early. However, a lot of these go undetected, resulting in many avoidable deaths. A report from 2009 estimates that, for the case of cancer in the United Kingdom alone, five to ten thousand deaths could be prevented yearly through early diagnosis [1]. Improving early diagnosis could thus beneficially affect patient outcomes, but is impeded by several factors, including cost and invasiveness of relevant diagnostic procedures. Thus, one of the aims of clinical diagnostic model research is to find diagnostic models with good predictive performance, using the cheapest and least invasive means possible. Examples of such research are the studies organized by the International Ovarian Tumor Analysis [2-5] and International Endometrial Tumor Analysis (IETA) [6] consortia, which investigate diagnostic models for ovarian and endometrial tumors, respectively. Figure 1 shows a typical clinical diagnostic model research workflow.

Figure 1. Typical workflow of clinical diagnostic model research. The Clinical Data Miner software framework improves support for the steps indicated in green. Support for steps marked in blue is planned for future work. (Abbreviations used: CRF=case report form; eCRF=electronic CRF; API=application programming interface.).



Software to Support Clinical Diagnostic Model Research Workflow

Several software packages exist to support the clinical diagnostic model research workflow. Electronic case report form (eCRF) systems, such as REDCap [7] or the open-source OpenClinica, enable the collection of patient data. Compared with paper-based data collection, such systems reduce data error rates [8], and, according to a costs simulation study, enable cost reductions between 49% and 62% [9]. As a result, their use has greatly increased over the past decades, with reports of 41% out of 259 Canadian trials using electronic data capture software [10] , and of 79.6% (417/524) of Hong Kong private physicians using electronic medical records [11] .

Meanwhile, mathematical packages such as R [12], Matlab [13], or WEKA [14,15] support data analysis. Their inclusion of machine-learning techniques enables the extraction of sophisticated diagnostic models from patient data, with high predictive performance.

However, several steps in the clinical diagnostic model research workflow introduce unnecessary complexity. Data have to be extracted from the eCRF system, and imported back into data analysis software. These steps may lead to conversion issues, requiring manual inspection of the result. Furthermore, any case report form (CRF) structure information is lost in the process. For data preprocessing transformations, such as the replacement of categorical variables with dummy variables [16] , the lack of CRF structure information requires either manual selection or the use of heuristics for determining which variables need to be transformed, both of which are prone to errors. Other transformations, such as dealing with structurally missing variables, can only be performed manually.

Moreover, the complexity of the data analysis step discourages intermediate assessments of predictive performance. As a result, clinical diagnostic model research usually relies on Monte Carlo simulations [17] or rules of thumb [18] for sample size requirements estimates. These may be both over and underestimated, leading to patient recruitment that is more expensive than needed, or to models with insufficient predictive performance, respectively.

We implemented the Clinical Data Miner (CDM) software framework [19] to support the studies organized by the IETA consortium [6] . In doing so, we aimed to create a generic, multi-centric platform that avoids the aforementioned inefficiencies, with a user interface that can be integrated in various computing environments, such as mobile phones or hospital information systems (HIS).

Methods

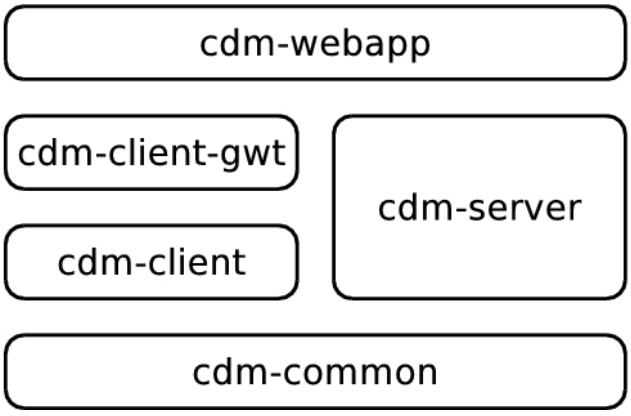
Component Overview

In order to improve support for clinical diagnostic model research in general, and the IETA studies in particular, the CDM software framework consists of an eCRF component and a data analysis component. This section introduces the eCRF and data analysis components in more detail, discusses the methodology used in their development, and explains the modalities of a survey we conducted to examine user satisfaction with CDM's eCRF component.

Electronic Case Report Form Component

CDM's eCRF component parses CRFs from external files, using a spreadsheet format similar to that of OpenClinica. Defining CRFs by parsing external files enables support for generic studies. In order to simplify the organization of multi-center studies, CDM's eCRF component exhibits a client-server architecture, with a Web-based user interface at the client side. This client-server architecture is reflected in the eCRF component's modular design. Figure 2 shows this, with separate modules for client and server code. The design further separates user interface *logic* (cdm-client) and user interface *presentation* (cdm-client-gwt). The latter separation offers the possibility to implement alternative interfaces, such as a mobile phone app, or a user interface integrated in a HIS.

Figure 2. In Clinical Data Miner (CDM)'s layered architecture, module cdm-common contains functionality common to client and server. The server code is implemented in module cdm-server, while client code is further split into user interface logic (cdm-client) and user interface presentation (cdm-client-gwt). Finally, cdm-webapp combines the modules and provides CDM's entry point.



Data Analysis Component

CDM includes capabilities for analyzing data, consisting of Java libraries for data querying and preprocessing, and the application of supervised machine-learning techniques. The simplified Unified Modeling Language diagrams from Figures 3 and 4 illustrate the application programming interfaces (APIs) of these libraries. Here, the DataManager class from Figure 3 represents CDM's entry point to its data querying and

preprocessing capabilities, while ClassifierFacade in Figure 4 provides access to its machine-learning capabilities.

The integration of an eCRF component with these data analysis libraries in a single system allows one to avoid exporting data from an eCRF system to import them back into data analysis software, eliminating potential conversion issues.

This integration additionally provides CDM's data preprocessing methods with direct access to CRF structure information. Instead of relying on manual input or heuristics, this direct access to CRF structure information enables preprocessing data with exact knowledge of type and dependency information for all variables. The createFactorProxies() preprocessor, for example, uses type knowledge of a CRF's variables to transform all categorical variables into sets of dummy variables [16]. Preprocessors such as flatten(), on the other hand, use information about dependencies between variables to convert data points with structurally missing variables to vectors. These are variables that may be missing depending on the value of a parent variable, as is the case for the variable “years past menopause” for patients with variable “menopausal status” set to “premenopausal”. By converting data points with structurally missing variables to vectors, the flatten() method enables the use of a wider variety of classification algorithms, such as logistic regression [20] or Least-Squares Support Vector Machines [21,22], without the need for defining specialized kernel methods.

Using the newWekaClassifier() method, the ClassifierFacade interface from Figure 4 constructs Classifier objects that provide access to the wealth of machine-learning algorithms and techniques available in the Weka toolbox [15]. Leveraging the Classifier interface, ClassifierFacade's sweep() method further enables the generation of learning curves, plotting the evolution of predictive performance measures, such as accuracy, sensitivity, specificity, or Area under the Receiver Operating Characteristic Curve, with respect to sample size.

Finally, CDM's Java libraries for data querying, data preprocessing, and machine-learning can be used interactively from within a Jython console by means of a set of Jython modules included in CDM.

Figure 3. The DataManager application programming interfaces includes methods to access and preprocess data.

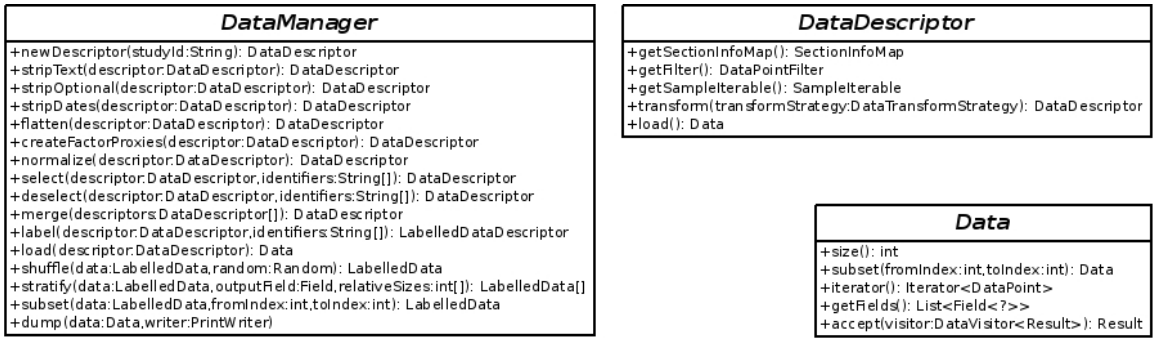
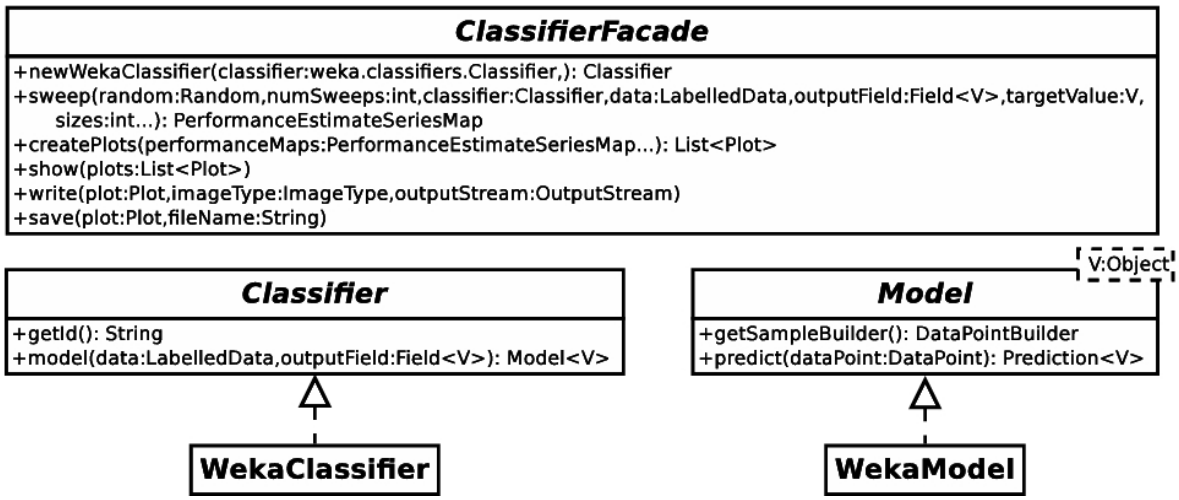


Figure 4. Unified Modeling Language diagram of Clinical Data Miner (CDM)'s machine-learning application programming interfaces. ClassifierFacade is the entry point to CDM's machine-learning functionality, which operates on Classifier objects to obtain Model objects.



Software Development Methodology

We developed CDM using the Java programming language, leveraging the Google Window Toolkit (GWT) to translate client-side Java code to ECMAScript. In order to ensure good software quality, we developed CDM using a test-driven development (TDD) [22] process. We have integrated Cobertura [23] in CDM’s automated build process for test coverage monitoring. The resulting unit test suite allows automation of most of the quality assurance process required prior to the deployment of new releases.

Sound design and loose coupling are obtained through extensive use of design patterns [23] and dependency injection. The latter

is achieved by means of the Spring framework server-side, and the Gin and Guice frameworks client-side.

User Survey

In order to assess user satisfaction, we sent a survey to active CDM users, which included users who submitted at least ten patient entries through CDM's eCRF component, or who participated in an interrater agreement study organized using CDM's user interface, adapted for such studies. In total, we asked 42 clinicians to participate in the survey. The survey consisted of several questions examining user-friendliness, satisfaction with certain user interface elements, and software reliability.

Results

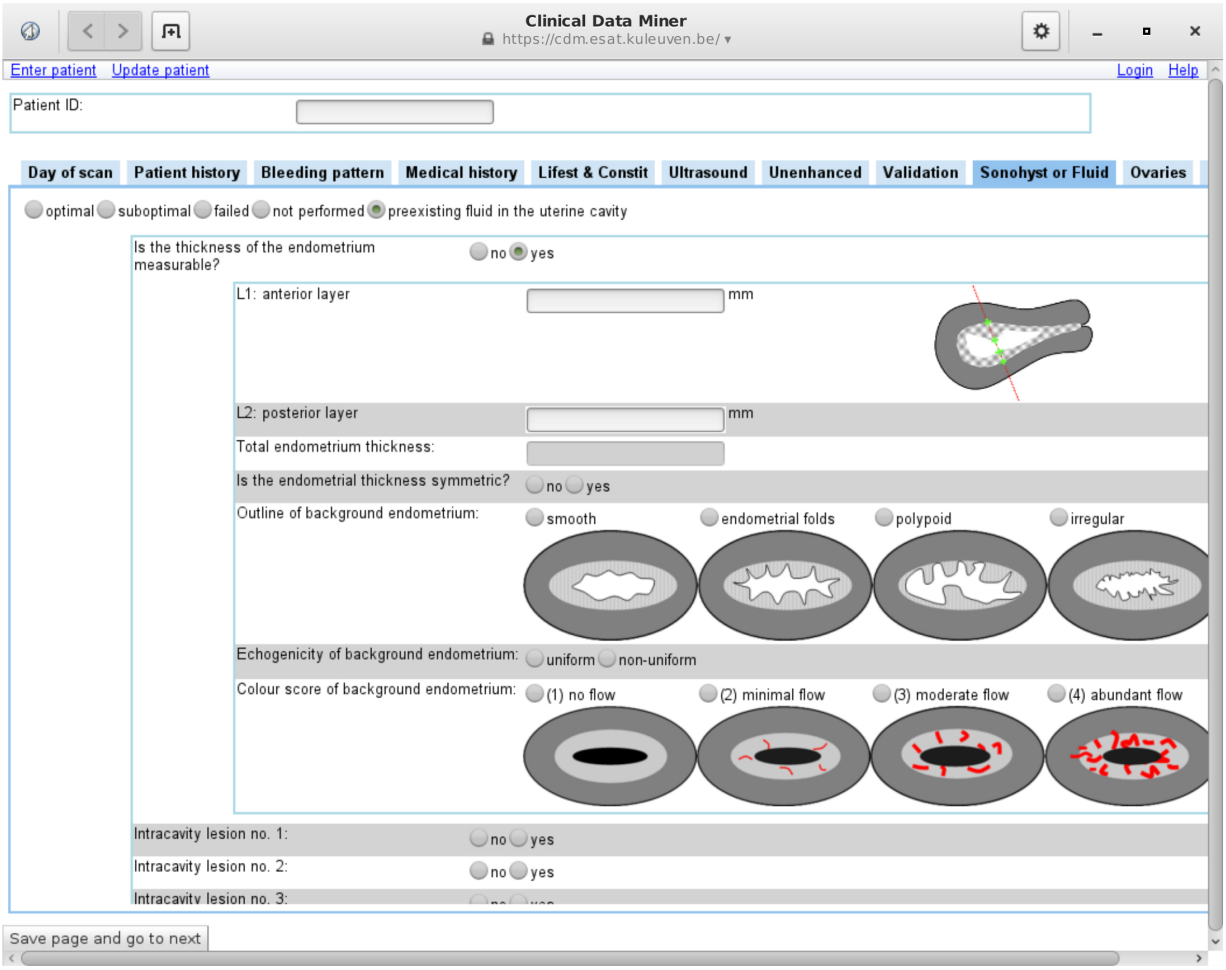
Electronic Case Report Form Component

CDM has a client-server architecture. As Figure 5 illustrates, its current user interface is Web-based. This has enabled multi-center data collection in the context of the IETA studies. As Table 1 shows, CDM has collected 4035 patient entries so far for these studies, supplied by 39 participants from 24 different centers between May 2011 and September 2014.

Table 1. Number of patient entries collected by CDM for the IETA studies, between May 2011 and September 2014.

IETA	Complete entries	Total entries
#1	1600	2069
#3	641	787
#4	891	1179
Total	3132	4035

Figure 5. Clinical Data Miner (CDM)'s data collection user interface. The possibility to include pictograms in case report forms is particularly interesting for variables obtained from imaging modalities.



Clinical Data Miner Architecture

CDM's modular, layered architecture enables parallel development of user interfaces for multiple computing environments, which in the future could thus include mobile phones or HIS. Moreover, the modularity of this architecture has facilitated the organization of interrater agreement studies that evaluate imaging modalities with the creation of a modified

user interface that displays each imaging modality next to the questionnaire to be completed.

Thanks to CDM's generic nature, other studies are planned for inclusion in CDM's eCRF system, such as studies about optimal cytoreduction, pregnancies of unknown location, and fertility. CDM's modified user interface for conducting interrater agreement studies has been used for the six studies listed in Table 2.

Table 2. List of interrater agreement studies organized using CDM user interface modified for supporting such studies.

	Study	Phases	Reference
1	Improvement of interrater agreement through pictograms	With and without pictograms	[24], [15]
2	Endo-myometrial junction	1 and 2	[25-27], [16-18]
3	Polycystic ovaries	1, 2a, 2b	[28], [19]
4	Uterine anomalies	-	[29], [20]
5	IETA 2	-	Not yet published[a]
6	Contrast enhancement study	with and without enhanced contrast images	Not yet published[b]

[a]Data were collected between July 2012 and February 2013. Authors: L Valentin, A Installé, P Sladkevicius, D Timmerman, B Benacerraf, L Jokubkiene, A diLegge, A Votino, L Zannoni, and T Van den Bosch.

[b]Data were collected between May 2013 and February 2014. Authors: A Sayasneh, A Installé, D Timmerman, T Van den Bosch, T Bourne, S Guerriero, F Rizzello, LPG Francesco, MA Pascual, A Rossi, A Czekierdowski, A Testa, E Coccia, and A Smith.

Data Analysis Component

CDM's data analysis APIs, and its Jython modules in particular, considerably simplify the derivation of machine-learning models from patient data. The integration of these capabilities into an eCRF system simplifies access to data, and the availability of the CRF definition simplifies preprocessing. Combined with the possibility to use these APIs interactively, CDM provides an excellent platform for rapid experimentation with different combinations of preprocessors and machine-learning algorithms in order to examine which combinations optimize predictive performance.

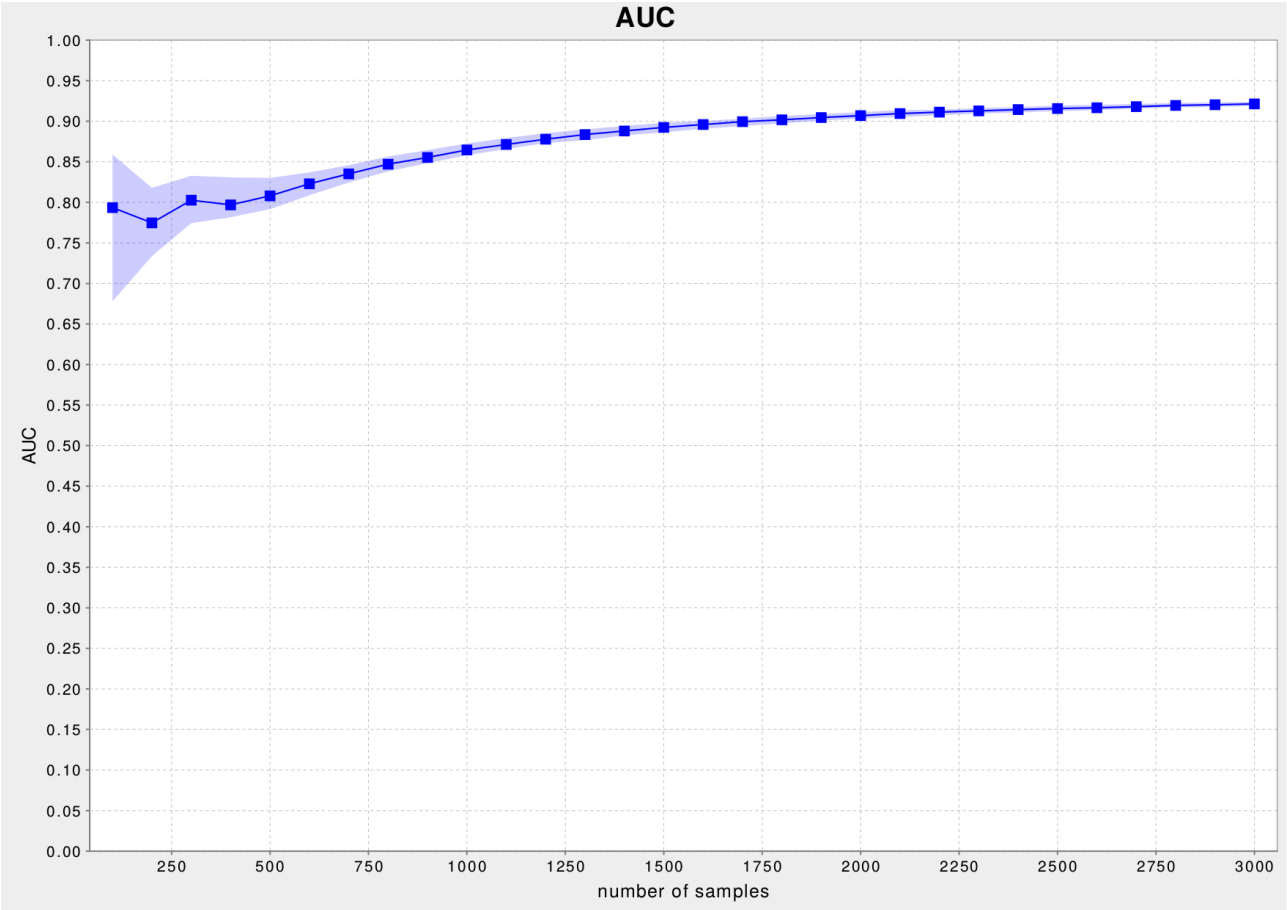
CDM's APIs provide a method for easily generating learning curves; Figure 6 shows one of these. Such curves offer a clear insight into the evolution of a study's predictive performance as the number of patient inclusions grows, so that study coordinators can make an informed decision whether to continue or to terminate enrolling patients. As long as growing patient numbers result in marked performance improvements, patient data collection should continue in order to generate better models. By contrast, if the learning curves hit a plateau, or exhibit a slope that is negligible with respect to variability of performance results, patient recruitment should be terminated in order to avoid useless patient recruitment costs. The ability to optimize costs associated with patient enrollment results in more optimal patient numbers than Monte Carlo simulations [17] or rules of thumb [18] could provide, and has been very well received by the IETA consortium's steering committee.

Table 3. Breakdown per module of number of source lines of code (SLOC) and line and branch test coverage ratios, as determined by the sloccount and Cobertura programs, respectively.

	Production code (SLOC[a])	Test code (SLOC)	Line coverage n (%)	Branch coverage n (%)
cdm-common	5862	7023	1800/1957 (91.98)	459/486 (94.4)
cdm-server	15,260	28,109	5781/6250 (92.50)	1437/1577 (91.12)
cdm-client	3595	7607	1128/1269 (88.89)	133/146 (91.1)
cdm-client-gwt	4090	5123	957/1828 (52.35)	137/321 (42.7)
cdm-webapp	321	177	38/111 (34.2)	2/2 (100)
Total	29,128	48,039	-	-
Weighted average	-	-	9704/11,415 (85.01)	2168/2532 (85.62)

[a]Note that interfaces contribute to SLOC, but not to the number of lines analyzed for line coverage, leading to different counts for number of lines in the “Production code” and “Line coverage” columns.

Figure 6. Learning curves, plotting predictive performance with respect to number of patient inclusions, can easily be generated using Clinical Data Miner (CDM)'s libraries. (Abbreviations: AUC=area under the ROC curve; ROC=receiver operating characteristic.).



Software Development Methodology

Our TDD approach has delivered good test coverage, as is apparent from Table 3. Modules cdm-common, cdm-server, and cdm-client all have line and branch test coverage levels around 90%, guaranteeing high software quality. Modules cdm-client-gwt and cdm-webapp, responsible for binding graphical widgets to user interface logic, and therefore difficult to verify using unit tests, have lower test coverages. However, thanks to these latter modules' low complexity and infrequent changes, their lower test coverages do not negatively affect software quality.

Survey

Out of 42 clinicians contacted, 28 responded, resulting in a response rate of 67%. Survey results in Table 4 show CDM to be considered user friendly. Users particularly appreciate the possibility to integrate pictograms for clarifying questions. A large majority of users, 79% (22/28), experienced problems in less than 5% of interactions with CDM; Figure 7 shows this information. All respondents considered using CDM for the organization of their own studies.

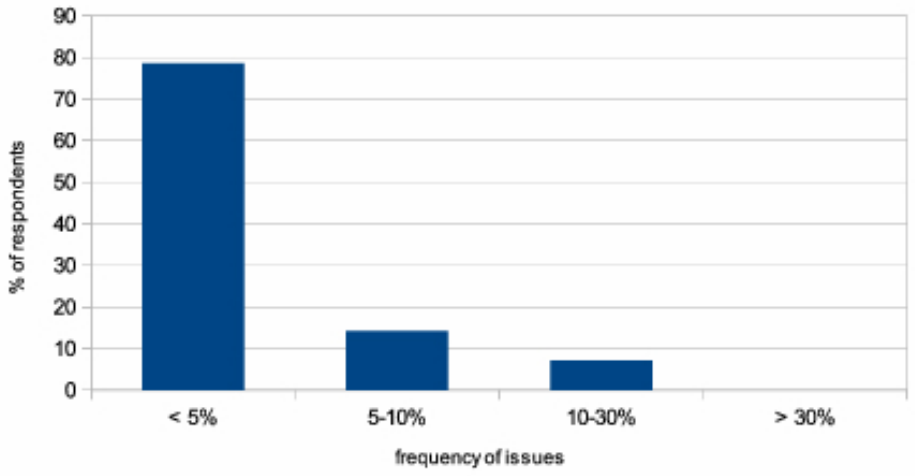
Table 4. Average agreement levels with survey propositions among respondents.

Proposition	Average agreement[a]
CDM is user-friendly.	8.6
The layout of studies is clear.	8.6
The VAS[b] is user-friendly.	8.1
CDM's VAS[b] is a good alternative to a paper VAS[b].	8.2
Pictograms help to clarify questions.	9.4
Pictograms help to differentiate multiple choice questions.	9.2
Pictograms next to multiple choice options will improve reliability.	9.3

[a]0 = no agreement; 10 = full agreement

[b]VAS = visual analog scale

Figure 7. Distribution of respondents over different ranges of issue frequencies. A large majority, 79% (22/28), of survey participants experienced problems in less than 5% of their interactions with Clinical Data Miner.



Discussion

Principal Findings

We developed an eCRF software framework for supporting generic, multi-center clinical studies. Its high test coverage guarantees good software quality and good maintainability, while its modular architecture ensures the framework's extensibility.

Its built-in data access, data preprocessing, and machine-learning capabilities streamline the clinical diagnostic model research workflow by eliminating data export and import steps, as well as by simplifying preprocessing. The possibility to access these capabilities through a Jython console provides an excellent platform for experimenting with different combinations of preprocessing and machine-learning algorithms.

The functionality to simplify the generation of learning curves enables study coordinators to assess whether to continue or to terminate data collection, providing better dataset size estimates than a priori application of rules of thumb or Monte Carlo simulations could deliver.

Limitations

CDM does not currently support variable length array types, reducing its usefulness for longitudinal data capture. For bounded array sizes, presenting a fixed amount of fields representing the array can alleviate this issue.

CDM's data analysis capabilities are currently only accessible through a Java API or a Jython console, requiring programming expertise for their use.

Future work should solve these limitations, with better support for longitudinal data, and the integration of data analysis capabilities into CDM's user interface. The latter will, for example, enable study coordinators to visualize learning curves directly from within the user interface.

Conclusions

The integration of data collection, preprocessing, and machine-learning in a single software framework simplifies the diagnostic model research workflow. The functionality for generating learning curves enables study coordinators to improve dataset size requirement estimates, also improving efficiency of clinical diagnostic model research.

Acknowledgments

BDM and DT are full professors at KU Leuven, Belgium. DT is Senior Clinical Investigator of Research Foundation - Flanders (FWO). CDM is a project cofunded by iMinds, an independent research institute founded by the Flemish Government. Project partner is KU Leuven Department of Obstetrics and Gynecology, University of Leuven.

This research is supported by the Research Council KU Leuven, GOA/10/09 MaNet; FWO project G.0871.12N (Neural circuits); Agency for Innovation by Science and Technology (IWT); TBM-Logic Insulin(100793), TBM Rectal Cancer(100783), TBM IETA(130256); PhD grants; Industrial Research Fund (IOF), IOF/HB/13/027 Logic Insulin; iMinds Medical Information Technologies Strategic Basic Research (SBO) 2014; Flemish League against Cancer (VLK) E. van der Schueren Foundation, rectal cancer; Federal Government, Federal Public Service (FOD), Cancer Plan 2012-2015 KPC-29-023 (prostate); European Cooperation in Science and Technology (COST), action BM1104, Mass Spectrometry Imaging.

The scientific responsibility is assumed by its authors.

Conflicts of Interest

None declared.

References

1. Richards MA. The size of the prize for earlier diagnosis of cancer in England. Br J Cancer 2009 Dec 3;101 Suppl 2:S125-S129 [FREE Full text] [doi: 10.1038/sj.bjc.6605402] [Medline: 19956156]
2. Timmerman D, Bourne T, Tailor A, Collins WP, Verrelst H, Vandenberghe K, et al. A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: The development of a new logistic regression model. Am J Obstet Gynecol 1999 Jul;181(1):57-65. [Medline: 10411796]
3. Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML, et al. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: A multicenter study by the International Ovarian Tumor Analysis Group. J Clin Oncol 2005 Dec 1;23(34):8794-8801 [FREE Full text] [doi: 10.1200/JCO.2005.01.7632] [Medline: 16314639]
4. Timmerman D, Testa AC, Bourne T, Ameye L, Jurkovic D, Van Holsbeke C, et al. Simple ultrasound-based rules for the diagnosis of ovarian cancer. Ultrasound Obstet Gynecol 2008 Jun;31(6):681-690 [FREE Full text] [doi: 10.1002/uog.5365] [Medline: 18504770]
5. Van Calster B, Valentin L, Van Holsbeke C, Testa AC, Bourne T, Van Huffel S, et al. Polytomous diagnosis of ovarian tumors as benign, borderline, primary invasive or metastatic: Development and validation of standard and kernel-based risk prediction models. BMC Med Res Methodol 2010;10:96 [FREE Full text] [doi: 10.1186/1471-2288-10-96] [Medline: 20961457]
6. Leone FPG, Timmerman D, Bourne T, Valentin L, Epstein E, Goldstein SR, et al. Terms, definitions and measurements to describe the sonographic features of the endometrium and intrauterine lesions: A consensus opinion from the International Endometrial Tumor Analysis (IETA) group. Ultrasound Obstet Gynecol 2010 Jan;35(1):103-112 [FREE Full text] [doi: 10.1002/uog.7487] [Medline: 20014360]
7. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
8. Walther B, Hossin S, Townend J, Abernethy N, Parker D, Jeffries D. Comparison of electronic data capture (EDC) with the standard data capture method for clinical trial data. PLoS One 2011;6(9):e25348 [FREE Full text] [doi: 10.1371/journal.pone.0025348] [Medline: 21966505]
9. Pavlović I, Kern T, Miklavcic D. Comparison of paper-based and electronic data collection process in clinical trials: Costs simulation study. Contemp Clin Trials 2009 Jul;30(4):300-316. [doi: 10.1016/j.cct.2009.03.008] [Medline: 19345286]
10. El Emam K, Jonker E, Sampson M, Krleza-Jerić K, Neisa A. The use of electronic data capture tools in clinical trials: Web-survey of 259 Canadian trials. J Med Internet Res 2009;11(1):e8 [FREE Full text] [doi: 10.2196/jmir.1120] [Medline: 19275984]
11. Cheung CS, Tong EL, Cheung NT, Chan WM, Wang HH, Kwan MW, et al. Factors associated with adoption of the electronic health record system among primary care physicians. JMIR Med Inform 2013 Aug 26;1(1):e1. [doi: 10.2196/medinform.2766]
12. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. URL: http://web.mit.edu/r_v3.0.1/fullrefman.pdf [accessed 2014-09-18] [WebCite Cache ID 6Sgera9CL]
13. The Mathworks, Inc. MATLAB and statistics toolbox release 2010b. Natick, Massachussets, United States: The Mathworks, Inc; 2010. URL: http://www.walkingrandomly.com/?P=4767 [accessed 2014-09-18] [WebCite Cache ID 6Sgf1UtjM]
14. Witten IH, Frank E. Data mining: Practical machine learning tools and techniques. Amsterdam: Morgan Kaufman; 2005.
15. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: An update. SIGKDD Explor. Newsl 2009 Nov 16;11(1):10. [doi: 10.1145/1656274.1656278]
16. Suits DB. Use of dummy variables in regression equations. Journal of the American Statistical Association 1957 Dec;52(280):548-551. [doi: 10.2307/2281705]
17. Muthén LK, Muthén BO. How to use a Monte Carlo study to decide on sample size and determine power. Structural Equation Modeling: A Multidisciplinary Journal 2002 Oct;9(4):599-620. [doi: 10.1207/S15328007SEM0904_8]
18. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996 Dec;49(12):1373-1379. [Medline: 8970487]
19. Installé AJF. Clinical Data Miner – Towards more efficient clinical study support. Leuven, Belgium: KU Leuven; 2014 Jun. URL: https://lirias.kuleuven.be/bitstream/123456789/452854/1/thesis.pdf [accessed 2014-09-18] [WebCite Cache ID 6TLsfb9oT]
20. Hosmer DW, Lemeshow S. Applied logistic regression. New York: Wiley; 2000.
21. Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett 1999;9(3):293-300. [doi: 10.1023/A:1018628609742]

22. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Least squares support vector machines. River Edge, NJ: World Scientific; 2002.
23. Gamma E. Design patterns: Elements of reusable object-oriented software. Reading, Mass: Addison-Wesley; 1995.
24. Installé AJF, Van den Bosch T, Van Schoubroeck D, Heymans J, Zannoni L, Jokubkiene L, et al. Showing pictograms in electronic data capture software improves inter-rater agreement. In: Ultrasound Obstet Gynecol.: Wiley & Sons Ltd; 2011 Presented at: Proceedings of the 21st World Congress in Obstetrics & Gynecology; Sept 2011; Los Angeles, USA p. 18-22 URL: http://onlinelibrary.wiley.com/doi/10.1002/uog.9334/full [doi: 10.1002/uog.9334]
25. Votino A, Installé AJF, Van den Bosch T, Van Schoubroeck D, Kacem Y, Kaijser J, et al. Optimal ultrasound visualization of the endometrial-myometrial junction (EMJ). In: Ultrasound Obstet Gynecol.: Wiley & Sons Ltd; 2012 Sep Presented at: Proceedings of the 22nd World Congress in Obstetrics and Gynecology; Sept 2012; Copenhagen, Denmark p. 9-12 URL: http://onlinelibrary.wiley.com/doi/10.1002/uog.11748/full
26. Votino A, Installé AJF, Van Pachterbeke C, Van Schoubroeck D, Kacem Y, Kaijser J, et al. Optimization of the image quality of endometrial-myometrial junction (EMJ). In: Ultrasound Obstet Gynecol.: Wiley & Sons Ltd; 2012 Sep Presented at: Proceedings of the 22nd World Congress in Obstetrics and Gynecology; Sept 2012; Copenhagen, Denmark p. 9-12 URL: http://onlinelibrary.wiley.com/enhanced/doi/10.1002/uog.11747/
27. Votino A, Installé AJF, Van den Bosch T, Van Schoubroeck D, Kacem Y, Kaijser J, et al. The influence of patient characteristics on the image quality of the endometrial-myometrial junction (EMJ). In: Ultrasound Obstet Gynecol.: Wiley & Sons Ltd; 2012 Sep Presented at: Proceedings of the 22nd World Congress in Obstetrics and Gynecology; Sept 2012; Copenhagen, Denmark p. 9-12 URL: http://onlinelibrary.wiley.com/doi/10.1002/uog.11413/full
28. Van Schoubroeck D, Installé AJF, Raine-Fenning NJ, De Neubourg D, Van den Bosch T, De Moor B, et al. Interobserver variability in the ultrasound diagnosis of polycystic ovaries using pattern recognition. In: Ultrasound Obstet Gynecol.: Wiley & Sons Ltd; 2012 Sep Presented at: Proceedings of the 22nd World Congress in Obstetrics and Gynecology; Sept 2012; Copenhagen, Denmark p. 9-12 URL: http://onlinelibrary.wiley.com/doi/10.1002/uog.11493/full
29. Van Schoubroeck D, Installé AJF, Raine-Fenning NJ, De Neubourg D, Van den Bosch T, De Moor B, et al. Interobserver variability in the ultrasound diagnosis of congenital uterine anomalies. In: Ultrasound Obstet Gynecol.: Wiley & Sons Ltd; 2012 Sep Presented at: Proceedings of the 22nd World Congress in Obstetrics and Gynecology; Sept 2012; Copenhagen, Denmark p. 9-12 URL: http://onlinelibrary.wiley.com/doi/10.1002/uog.11490/full

Abbreviations

API: application programming interface
CDM: Clinical Data Miner
COST: European Cooperation in Science and Technology
CRF: case report form
eCRF: electronic case report form
FOD: Federal Public Service
FWO: Research Foundation - Flanders
HIS: hospital information system
IETA: International Endometrial Tumor Analysis
IOF: Industrial Research Fund
IWT: Agency for Innovation by Science and Technology
SBO: Strategic Basic Research
SLOC: source lines of code
TBM: Applied Biomedical Research
TDD: test-driven development
VLK: Flemish League against Cancer

Edited by G Eysenbach; submitted 15.01.14; peer-reviewed by CH Li, H Zhai; comments to author 03.02.14; revised version received 18.07.14; accepted 17.08.14; published 20.10.14.

Please cite as:
Installé AJF, Van den Bosch T, De Moor B, Timmerman D
Clinical Data Miner: An Electronic Case Report Form System With Integrated Data Preprocessing and Machine-Learning Libraries Supporting Clinical Diagnostic Model Research
JMIR Med Inform 2014;2(2):e28
URL: http://medinform.jmir.org/2014/2/e28/
doi:10.2196/medinform.3251
PMID:25600863

Original Paper

CohortExplorer: A Generic Application Programming Interface for Entity Attribute Value Database Schemas

Abhishek Dixit[1], BSc (Hons), MRes; Richard J B Dobson[1], BSc (Hons), PhD (bioinformatics)

Institute of Psychiatry, NIHR Biomedical Research Centre for Mental Health & Biomedical Research Unit for Dementia, South London and Maudsley NHS Foundation Trust & Institute of Psychiatry, Kings College London, London, United Kingdom

Corresponding Author:
Abhishek Dixit, BSc (Hons), MRes
Institute of Psychiatry
NIHR Biomedical Research Centre for Mental Health & Biomedical Research Unit for Dementia
South London and Maudsley NHS Foundation Trust & Institute of Psychiatry, Kings College London
De Crespigny Park
London, SE5 8AF
United Kingdom
Phone: 44 20 7848 0924
Fax: 44 20 7848 0866
Email: abhishek.dixit@kcl.ac.uk

Abstract

Background: Most electronic data capture (EDC) and electronic data management (EDM) systems developed to collect and store clinical data from participants recruited into studies are based on generic entity-attribute-value (EAV) database schemas which enable rapid and flexible deployment in a range of study designs. The drawback to such schemas is that they are cumbersome to query with structured query language (SQL). The problem increases when researchers involved in multiple studies use multiple electronic data capture and management systems each with variation on the EAV schema.

Objective: The aim of this study is to develop a generic application which allows easy and rapid exploration of data and metadata stored under EAV schemas that are organized into a survey format (questionnaires/events, questions, values), in other words, the Clinical Data Interchange Standards Consortium (CDISC) Observational Data Model (ODM).

Methods: CohortExplorer is written in Perl programming language and uses the concept of SQL abstract which allows the SQL query to be treated like a hash (key-value pairs).

Results: We have developed a tool, CohortExplorer, which once configured for a EAV system will "plug-n-play" with EAV schemas, enabling the easy construction of complex queries through an abstracted interface. To demonstrate the utility of the CohortExplorer system, we show how it can be used with the popular EAV based frameworks; Opal (OBiBa) and REDCap.

Conclusions: The application is available under a GPL-3+ license at the CPAN website. Currently the application only provides datasource application programming interfaces (APIs) for Opal and REDCap. In the future the application will be available with datasource APIs for all major electronic data capture and management systems such as OpenClinica and LabKey. At present the application is only compatible with EAV systems where the metadata is organized into surveys, questionnaires and events. Further work is needed to make the application compatible with EAV schemas where the metadata is organized into hierarchies such as Informatics for Integrating Biology & the Bedside (i2b2). A video tutorial demonstrating the application setup, datasource configuration, and search features is available on YouTube. The application source code is available at the GitHub website and the users are encouraged to suggest new features and contribute to the development of APIs for new EAV systems.

(JMIR Med Inform 2014;2(2):e32) doi:10.2196/medinform.3339

KEYWORDS

entity-attribute-value schema; biobank database; clinical information systems; CDISC ODM; SQL

Introduction

Electronic data capture (EDC) and electronic data management (EDM) systems are a key requirement for studies in modern biomedical science. Such systems are developed to centrally manage the recruitment and storage of participant details. They typically comprise a powerful database engine accessible over the network using Web-based technologies. Such systems

include built-in sanity checking and quality control procedures to ensure the data captured is consistent and well formatted for ease of downstream analysis. Some of the popular EDC and EDM systems include OpenClinica, LabKey, Onyx, Opal, REDCap, entity-attribute-value **(**EAV) with classes and relationships (EAV/CR), and Informatics for Integrating Biology & the Bedside (i2b2) [1-7].

Typically, EDC and EDM systems employ a generic EAV schema. Through the use of such a schema, hundreds of clinical attributes (or variables) can be stored in a single table without having to create multiple tables. Additionally, more attributes can be easily added without changing the underlying schema [8-11]. The EAV model can be viewed as a database table with three columns; one column specifies the entity (eg, participant ID), one for the attribute (eg, cognitive test), and one for the value of the attribute (eg, cognitive test score) [12]. If the study is longitudinal with multiple follow-up visits then an additional column is often used to store the visit number. In longitudinal studies, the combination of entity_id and visit number can act as a primary key (a composite primary key).

Although the EAV model provides a great deal of flexibility in storing data, such a schema requires the use of complex structured query language (SQL) to extract subsets of data from the tables [13-16]. In addition, the choice of the system depends on the study requirements. This poses a problem for the researchers as different EDC and EDM systems differ in their graphical user interface, data model and sometimes the vendor/relational database management system (eg, OpenClinica can be implemented in Oracle, PostgreSQL, Labkey in PostgreSQL, REDCap, and Opal in MySQL) thereby increasing the burden of understanding and using the underlined data model before querying datasources.

To address this problem we have developed CohortExplorer, a generic framework that allows the detailed exploration of clinical data stored under the EAV schema which is organized into a survey format (questionnaires/events, questions, and values) using a standard search interface. The main objectives were to: (1) standardize the interface to EAV databases; (2) enable user-friendly querying of entities and variables (ie, meta data) at depth; and (3) provide the functionality to export the data which can be readily parsed and loaded by statistical software such as R for downstream analysis [17]. CohortExplorer has no schema and solely depends on the datasource API (discussed in the next section) and read-only connection made to the clinical repository implementing the EAV schema.

By way of example, we demonstrate the utility of CohortExplorer by connecting to and querying two commonly used EDC and EDM systems, namely Opal [4] and REDCap [5] both implementing their own version of the EAV schema in MySQL. Opal and REDCap greatly vary in their functionality; Opal is developed to manage the participants (ie, EDM) recruited as part of the clinical studies and relies on Onyx [3], its sister software to recruit the participants (ie, EDC). Both Onyx and Opal are developed as a part of OBiBa [18], a core project of the Population Project in Genomics Consortium (P3G), committed towards building high quality open source systems for biobanks. All OBiBa software along with their source code is available under the open source GPL3 license. REDCap encompasses both participant recruitment and management functionalities. REDCap was developed at the Vanderbilt University and is currently comprised of over 900 active institutional partners with bases all over the world. REDCap, unlike systems developed by OBiBa, is not open source but is available at no charge.

Methods

CohortExplorer Core Components and Implementation

CohortExplorer has three main components: (1) a datasource API and configuration file; (2) an SQL abstraction layer; and (3) a command line search interface. Both the SQL abstraction layer and command line query interface have been implemented using object oriented Perl [19] programming language. Data captured by the systems (questionnaires, surveys, and forms) are referred to herein as tables and the questions, which form part of the study, are termed variables with values being the answers to the questions.

First, the easy part of building an EAV-schema-agnostic API is achieving backend independence. CohortExplorer implements backend independence by the use of Perl module DBI [20]. DBI is independent of any database available in the backend and is responsible for taking all SQL commands and dispatching them to the appropriate driver for execution. Using CohortExplorer's datasource API (a Perl class) the users can define the entity, table, and variable structure under the EAV system. By structure we mean what database tables and columns are to be consulted to query data and meta data. The organization of entities, tables, and variables can be transformed into Perl hash (ie, data structure with key value pairs) using SQL::Abstract [21] discussed below. The Perl hash for entity, table and variable can vary with variation in EAV schema. In addition, the user authentication mechanism can also be defined in the datasource API. The datasource configuration file allows the user to define datasource settings including database connection details like dsn, username, and password (ie, it is the connector). The documentation detailing the API is available online [22]. A video tutorial aiming to give users an insight into application set-up including datasource configuration is also available online [23]. The tutorial with examples demonstrates various search features offered by the application.

Currently, CohortExplorer comes with built-in APIs for Opal and REDCap each catering to their own authentication mechanism and variation in the EAV schema. Therefore, the datasources stored within these systems can be queried using the current set-up (Figure 1).

We intend to provide APIs for other EDC and EDM systems such as LabKey [1] and OpenClinica [2] so the users can query the repositories implemented using these systems with same ease as Opal and REDCap. Opal and REDCap were the starting point considering their use at our institution. The application source code is available on GitHub, a popular platform for sharing and developing code [24]. The user community is

encouraged to contribute to the development of APIs appertaining to new EAV systems.

The security in CohortExplorer is implemented using the built-in authentication mechanism, setuid and Linux file permissions. The application runs under the taint mode which sets up special security checks including the check for unauthorized input. The security features ensure the user running the application has no access to the configuration files containing the connection details of the clinical repositories, the administrator is expected to create a read-only connection to the repository. Moreover, the application can be easily made to pay attention to user permission assigned within the repository. For example, the REDCap datasource API ensures only users who are allowed to export data in REDCap can use CohortExplorer. The API takes into account what variables and records are accessible to the user within REDCap. If some user is prohibited from viewing the identifiable information on participants in REDCap the API makes sure the user does not have access to the variables pertaining to the participant identifiable information (eg, participant's name, address, etc).

Second**,** at its core, CohortExplorer is powered by the SQL abstraction layer implemented using the Perl module, SQL::Abstract [21]. The abstraction layer serves two main purposes. Firstly, it allows SQL statements to be treated as a hash with SQL components (ie, -columns, -from, -where, -group_by, -order_by, and -having) as keys in the hash. The SQL statements to query data and meta data can easily be constructed from the entity, table and variable structures defined in the datasource subclass. As the EAV datasource can be cross-sectional or longitudinal, the second feature of the abstraction layer is that it enables the SQL generating engine to generalize the EAV schema as a 1 or 2 table database (static or dynamic) depending on the datasource type, hence making the easy and flexible construction of complex and dynamic SQL statements with placeholders. This is done to address the data heterogeneity at the forms/questionnaires/surveys level. The forms, which are only used once throughout the study, are grouped under static table (eg, participant demographics). This table is created by grouping or aggregating the form data on entity_id. Such table is applicable to cross-sectional studies but may also apply to longitudinal datasources. The forms which are used repeatedly throughout the study in the form of follow-up visits are grouped under dynamic table (eg, cognitive assessments). The dynamic table is created by grouping the data on the concerned forms on the entity_id and the visit number. Currently the application does not support querying datasources with multiple arms. In future the application may consider other table structures to address variation in data with respect to arms.

Third, the command line interface (CLI) is implemented using the Perl module CLI::Framework [25] (See Figure 2) and enables the user to query the clinical datasources. The CLI has two main components: (1) Application - this component authenticates the user and initializes CohortExplorer for the user specified datasource and dispatches the supplied command for further processing (See Figure 3); (2) Command - this component does the command specific processing and returns the output to the application component for display. The command component is divided into 5 main commands each of which performs a specific operation as described in Textbox 1.

Each command has a mandatory help section which details command usage with examples. CohortExplorer can also be run on the standard Linux shell so the user can easily set-up a report scheduling using the Linux in built Cron functionality.

Textbox 1. Command components.

1. **describe** - This command prints the datasource description in a tabular format where the table header is the entity count (ie, number of participants in the datasource/study) and table body contains the information appertaining to each table in the datasource. The first column in the table body is the table name (ie, questionnaire/surveys/forms) followed by table attributes (eg, variable count, associated label, etc) specified in the datasource API.
2. **find** - This command allows the user to find recorded variables using keywords which can be utilized to build an entity search query. The user can perform both case insensitive and fuzzy searches. The command prints the variable dictionary (ie, meta data) of variables meeting the search criteria in a tabular format where the first column is the name of the variable, second column is the table which records the variable and other columns include variable attributes (eg, variable type, categories, associated label, etc) specified in datasource API. The command looks for the presence of keywords in all variable attributes.
3. **search** - This command allows the user to search entities using variables of interest. The user can also impose conditions on variables using all valid SQL operators; =, !=, >, <, >=, <=, in, not_in, like, not_like, ilike, between, not_between, regexp, and not_regexp. The command includes auto-completion enabling the user to enter the first few characters of some command option/argument (eg, export directory, variable or table name) and press the completion key (ie, TAB) to fill-in the rest of the characters. At any time in CohortExplorer's console/interactive mode the user is able to view all tables and variables they have access to by simply pressing the TAB key. In addition, the command allows the user to view descriptive statistics and export data in csv format which can be easily parsed in statistical software like R for downstream analysis. The search command is available to both cross sectional and longitudinal datasources. When calculating descriptive statistics for variables belonging to dynamic tables the command groups the variables by visit. The command also includes a bookmarking feature which allows the user to save commands for future use.
4. **compare** - As the name suggests, the compare command allows the user to compare entities across visits. The command is only available to longitudinal datasources. The command allows the user to search and impose conditions at a visit level. Prefixes vAny, vLast, v1, v2, etc are added to variable names. For example: v1.var represents first visit of the variable 'var', v2.var represents second visit, vAny.var implies any visit, vLast.var last visit, and 'var' in this command simply represents all visits. The prefix vAny and vLast are abstract terms as vAny and vLast can be any visit (generally the last time a variable was recorded for some entity is not known in advance so practically any visit can be the last visit). The data exported via this command is formatted horizontally (ie, repeating variables) unlike the search command which exports the data vertically (ie, repeating entities) where each row represents an entity followed by the user provided visit variables (ie, dynamic table) or simply variables in case of static tables. The statistics produced in this command are calculated with respect to the entity_id and the number of observations for each variable is equivalent to the number of times or visits each variable was recorded for each entity.
5. **history** - The user can keep track of their previously saved commands using the history command. By specifying the show option the user can view all their saved commands along with the date-time stamp. The user can re-run any of the previously saved command or use the information in the commands (ie, options arguments) to build new commands.

Figure 1. Using the datasource API, CohortExplorer, via the authentication mechanism, is able to connect to secure EAV clinical repositories such as REDCap and Opal. The connection made to the clinical repository is expected to be read-only. Once connected, the user can explore the datasources stored within the repositories using the abstracted command line search interface.

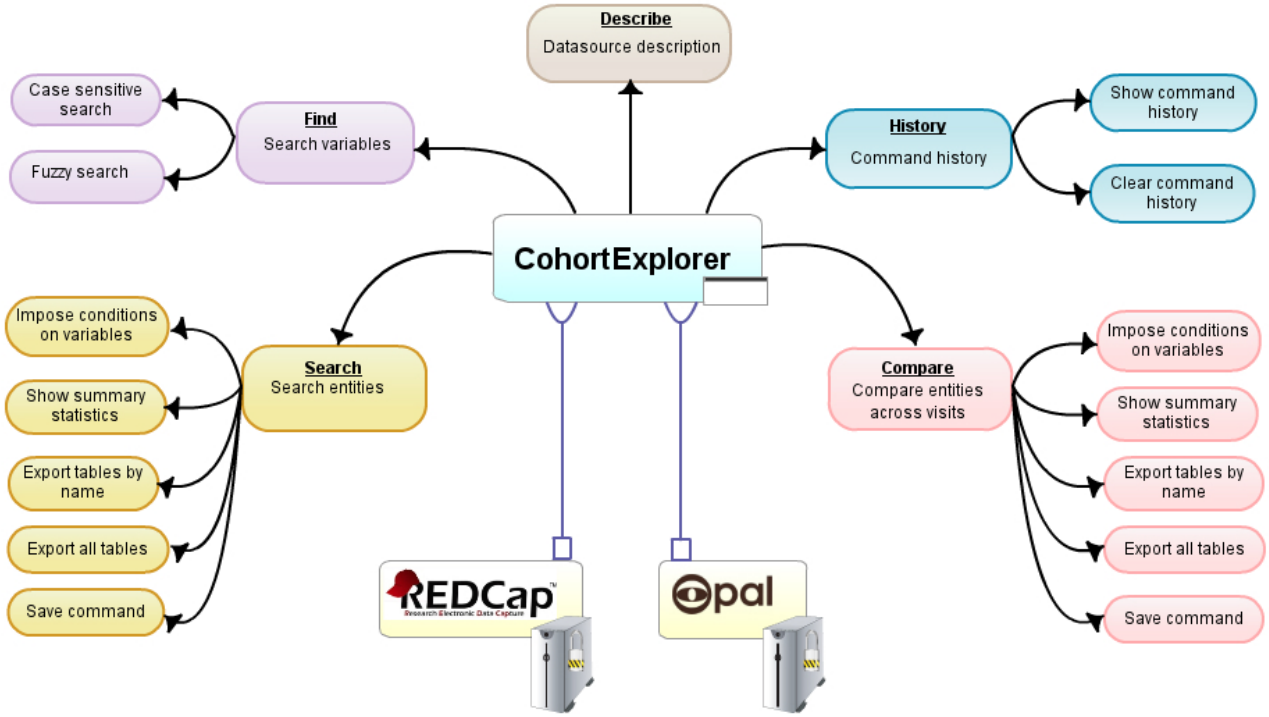


Figure 2. The figure shows the classes implemented under CohortExplorer. CohortExplorer::Application is inherited from CLI::Framework::Application (super class) and overrides the methods shown. All command classes are inherited from CLI::Framework::Command which is an abstract factory (dotted arrows). The meta commands such as menu and help are application aware commands with direct access to the application object. The application class dispatches the command objects. CohortExplorer::Command::Query extends CLI::Framework::Command (dark head arrow) and acts as an interface to search and compare command classes. CohortExplorer::Application class uses CohortExplorer::Datasource (an abstract class) to initialize the datasource. CohortExplorer::Datasource provides hooks or methods (also shown) which are implemented by subclasses such as CohortExplorer::Application::My::Datasource corresponding to electronic data capture and management systems such as Opal and REDCap.

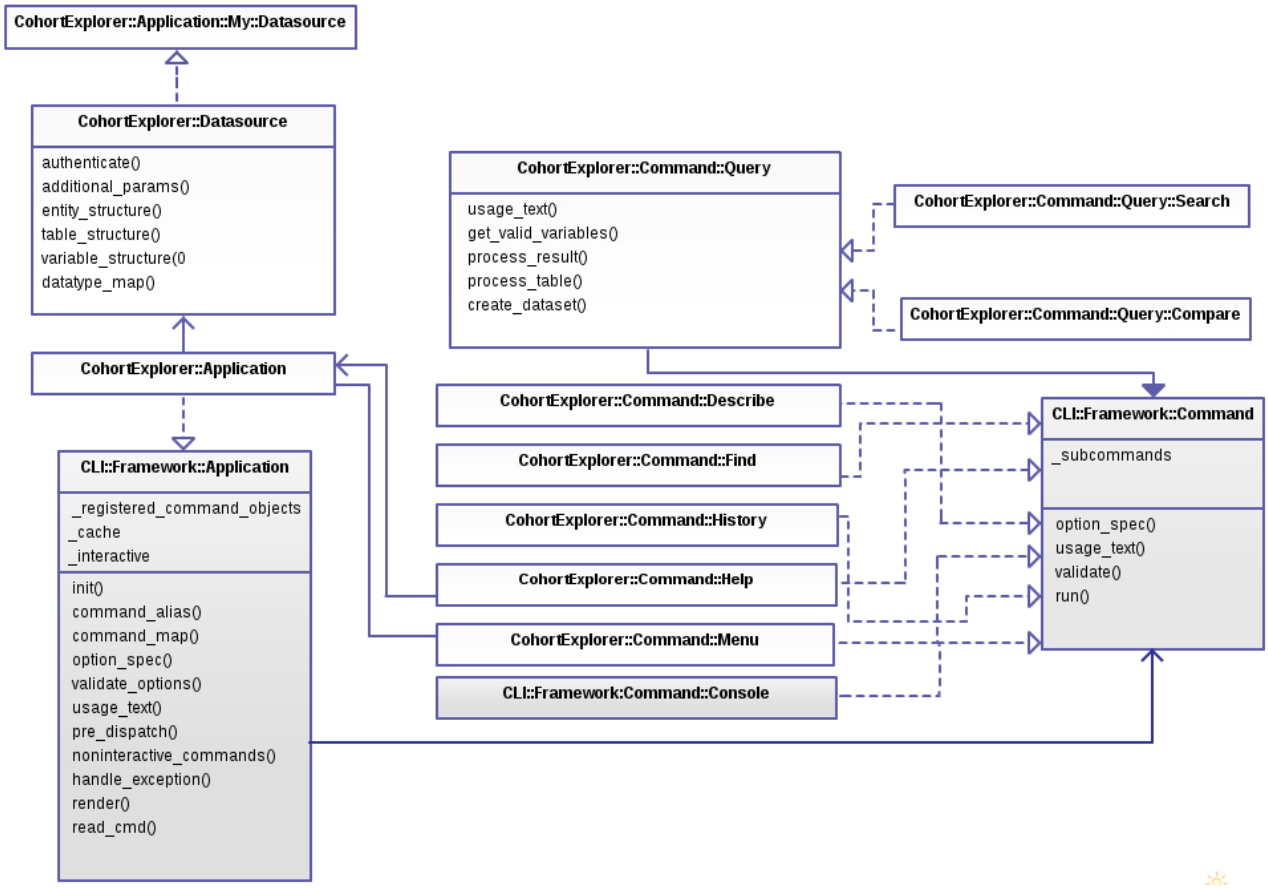
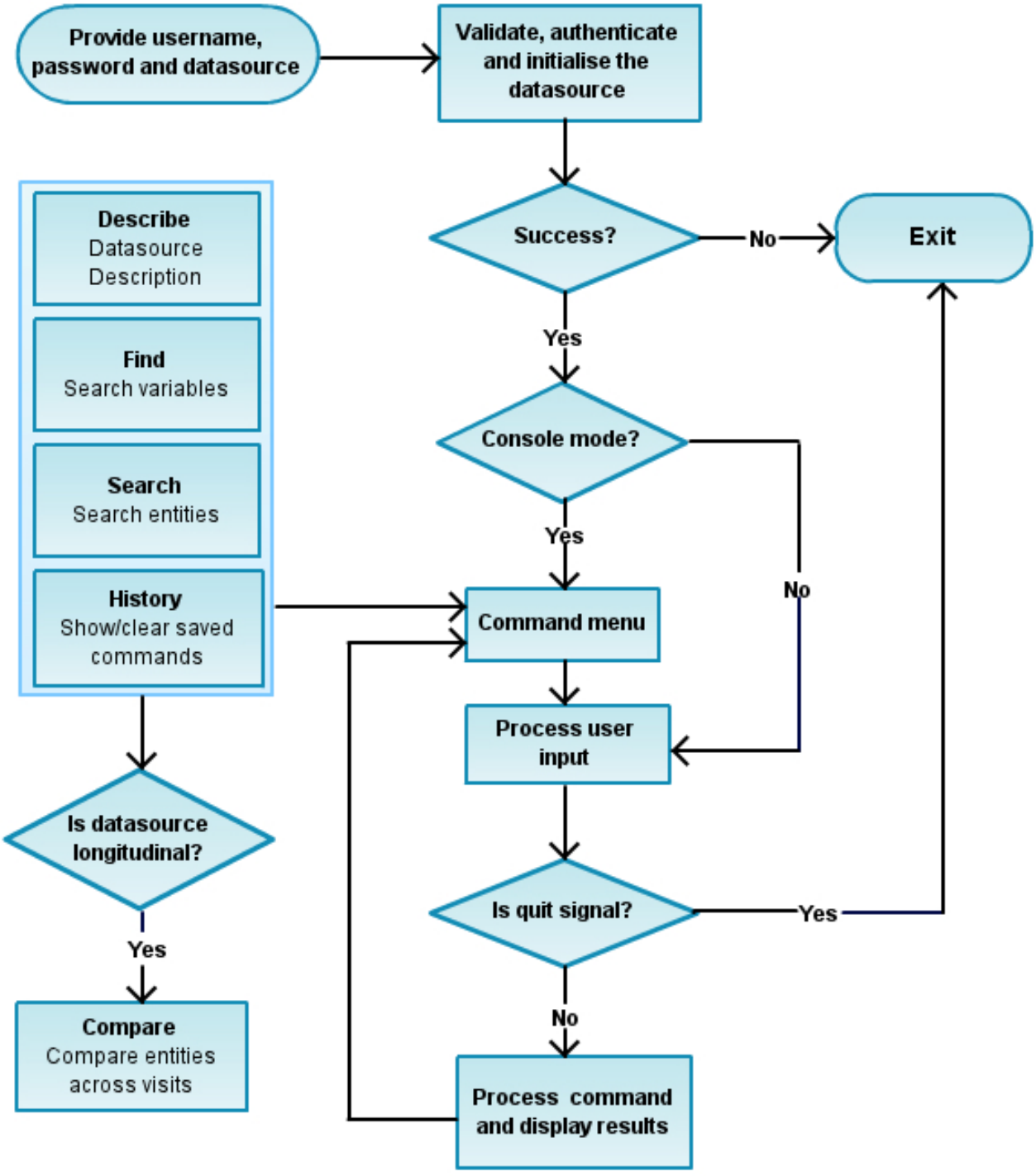


Figure 3. The flow chart showing the overall work flow of CohortExplorer. The user runs CohortExplorer by providing the datasource name, username, password and the command to execute. When running the application in console/interactive mode (determined by the console command) a menu of available commands is displayed. The user inputs the command along with command options and arguments (if applicable), the application processes the command and displays the results. When running the application on standard Linux shell the application simply processes the command input, displays the results and terminates the connection.



Results

Case Studies

Overview

To evaluate the utility and feasibility of CohortExplorer we connected the system to two real-world datasources, an Alzheimer's and Dementia biomarker study and the National Institute for Health Research **(**NIHR) BioResource for Mental Health based at the Institute of Psychiatry, Kings College London, United Kingdom.

The NIHR Alzheimer's and Dementia datasource is powered by Opal (OBiBa) [4] and stores the data from two cohorts; namely AddNeuroMed (European Union funded European Middleware Initiative) [26-27] and Kings Health Partners - Dementia Case Register (DCR) [28-29]. The data comprises participant and informant interviews conducted using Onyx (OBiBa) [3] longitudinally. The Onyx interview is comprised of7 main questionnaire categories: consent, demographics, physical measurement and samples obtained, disease history, family history, cognitive tests, and diagnosis.

Data collected by The NIHR BioResource for Mental Health [30] is stored in REDCap. This longitudinal study aims to collect 50,000 samples over the next 5 years from patients registered with South London And Maudsley NHS Foundation Trust (SLAM) and King's Health Partners. The project collects data on patients' demographics (eg, age, sex, ethnicity, etc) along with blood and saliva samples for molecular analysis, which includes developing new diagnosis tests, identifying new drug targets, and understanding the causes of different mental disorders.

Below, we provide examples of distributed queries that can easily be performed on the two clinical datasources using CohortExplorer's Opal and REDCap datasource API.

Alzheimer's and Dementia Datasource (Opal)

Questions we can answer using CohortExplorer's Opal datasource API (Figure 4): (1) during the course of the study how many participants with Mini Mental State Examination (MMSE) scores between 15 and 20 have had a history of hallucinations but not delusions or vice versa? We would like to know their disease status; (2) how many participants who previously had mild cognitive impairment have been diagnosed with Alzheimer's disease? We would also like to see their MMSE total at first and last visit; (3) at any visit during the study how many non-European females receiving anti-psychotic medication have been diagnosed with Alzheimer's disease? We would also like to know their MMSE scores and if they had ever suffered with high blood pressure and diabetes; and (4) how many participants have consented for brains for dementia research study? For all consented participants export complete data and show ethnicity, disease status at first and last visit.

Figure 4. CohortExplorer commands for the Alzheimer's and Dementia datasource. For better understanding the commands are divided into their respective components namely; command name, command options and command arguments.

cmd	[cmd-opts]	[cmd-args]
search	--out=/home/user/exports --stats --cond=MMSE.MMSE_Total='between, 15, 20' --cond=NPI.Hallucinations_Suffered='!=, NPI.Delusions_Suffered'	Conclusion.Disease_Status NPI.Delusions_Suffered
compare	--out=/home/user/exports --cond=vAny.Conclusion.Disease_Status='=, MCI' --cond=vLast.Conclusion.Disease_Status='=, ADC'	v1.MMSE.MMSE_Total vLast.MMSE.MMSE_Total
compare	--out=/home/user/exports --save-command --cond=FamilyHistory.Subject_Ethnicity='regexp, ^(white|british|europe)' --cond=vAny.Camdex.Antipsychotics='=, Yes' --cond=vAny.Camdex.Past_History_1='=, Yes' --cond=vAny.Camdex.Past_History_14='=, Yes' --cond=vLast.Conclusion.Disease_Status='=, ADC'	v1.MMSE.MMSE_Total vLast.MMSE.MMSE_Total Demographics.Education_Years
compare	--out=/home/user/exports --export-all --cond=v1.InformedConsent.Study_BDR='=, true' --cond=vLast.Conclusion.No_Contact_Permission_Reason='like, %died%'	v1.Conclusion.Disease_Status vLast.Conclusion.Disease_Status FamilyHistory.Subject_Ethnicity

NIHR BioResource Datasource (REDCap)

Questions we can answer using CohortExplorer's REDCap datasource API (Figure 5): (1) how many participants in the study were born between 1950 and 1970? For all consented participants, produce summary statistics showing percentage breakdown by gender and registration clinic; (2) how many females have withdrawn from the study citing negative media reports and health reasons? Show date of birth of all females; (3) how many participants have donated blood on their first visit but not the last visit? For all participants meeting the query criteria obtain data on gender, date of birth, and samples collected; and (4) how many participants have donated blood platelets in all visits? For all resulting participants show gender, date of birth, and the investigator who took the consent.

Figure 5. CohortExplorer commands for the NIHR BioResource for Mental Health datasource.

```
    cmd     |            [cmd-opts]                                                                  | [cmd-args]
------------|----------------------------------------------------------------------------------------|-----------------------------------
    search  |  --out=/home/user/exports                                                              | registration.gender
            |  --stats                                                                               | registration.registration_clinic
            |  --cond=registration.date_of_birth='between, 1950-01-01, 1970-01-01'                   |
            |  --cond=consent.consent_complete='=, 2'                                                |
            |                                                                                        |
------------|----------------------------------------------------------------------------------------|-----------------------------------
    search  |  --out=/home/user/exports                                                              | registration.date_of_birth
            |  --export=visit                                                                        | withdrawal.withdraw_instigator
            |  --cond=registration.gender='=, female'                                                |
            |  --cond=withdrawal.withdraw_reason='like, %negative%, %unwell%'                        |
            |                                                                                        |
------------|----------------------------------------------------------------------------------------|-----------------------------------
   compare  |  --out=/home/user/exports                                                              | registration.gender
            |  --save-command                                                                        | registration.date_of_birth
            |  --cond=v1.visit.visit_blood_donor='=, yes'                                            | v1.visit.visit_blood_samples
            |  --cond=vLast.visit.visit_blood_donor='=, no'                                          | v1.visit.visit_urine_samples
            |                                                                                        | vLast.visit.visit_blood_samples
            |                                                                                        | vLast.visit.visit_urine_samples
            |                                                                                        |
------------|----------------------------------------------------------------------------------------|-----------------------------------
   compare  |  --out=/home/user/exports                                                              | registration.gender
            |  --cond=visit.visit_platelets_donor='=, yes'                                           | registration.date_of_birth
            |                                                                                        | consent.consent_researcher
            |                                                                                        |
------------|----------------------------------------------------------------------------------------|-----------------------------------
```

Discussion

Principal Findings

CohortExplorer provides a secure and standard platform with which to query clinical repositories that are based on the EAV framework such as Opal and REDCap. The application relies on a read only database connection to the repository (via the datasource configuration file and datasource API). For longitudinal studies, CohortExplorer provides summary statistics both at the visit level as well as at the entity level. Moreover, the output from the application can be easily parsed by statistical software such as R for downstream analysis and the commands can be bookmarked for future use. The application runs on standard Linux shell thus making scheduled reports possible through the cron daemon. The application also supports the auto-complete or tab completion functionality making it easier for the user to provide variables and table names. The functionality can be helpful considering clinical variables can have long names.

CohortExplorer provided basic authorization and pays attention to the user permissions as implemented by the parent repository.

One of the main advantages of CohortExplorer is that the search interface is independent of the system storing the clinical data. This feature is of particular importance considering most of the EDC and EDM systems differ significantly in their query interface and researchers involved in multiple studies end up using multiple systems based on the study requirements. Deployment of CohortExplorer will lower the burden on researchers and data managers to learn and use the underlining data model before querying for entities of interest. With minimal training the researchers and data managers can use CohortExplorer to generate hypotheses, reports, and also to test the data accuracy.

CohortExplorer is written in Perl with CLI::Framework and SQL::Abstract as main modules. The application can be installed with all of its dependencies and the user manual via its Debian package which is available online [31]. As the application implements SQL abstraction it is compatible with other relational database management systems such as Oracle, Microsoft SQL Server, and PostgreSQL. However, this feature is yet to be tested. The Debian package includes Opal and REDCap APIs. The user is encouraged to use these as examples when trying to create a datasource API for a new EAV schema. The application is supported by active development and users are encouraged to suggest new features and get involved in development on GitHub [24]. At present, the application is only compatible with EAV systems that fit into a survey format (questionnaires/events, questions, and values) in other words, the CDISC Observational Data Model (ODM). Further work is needed to make the application compatible with EAV schemas where the metadata is organized into hierarchies such as i2b2.

The future work also includes extending the application to EDC and EDM systems implemented in Oracle, PostgreSQL, and Microsoft SQL server such as LabKey and OpenClinica.

Conclusions

CohortExplorer provides a user-friendly and generic approach to slice and dice clinical datasources stored under the EAV format. For biomedical researchers, CohortExplorer provides an easy to understand view of the unstructured and complex clinical data. The application is available as open source under the GPL-3+ license. The source-code, Debian package and manual are available online [24,31]. A video tutorial demonstrating the application set-up and features is also available online [23]. The tutorial aims to give users an insight into the application set-up, datasource configuration, and query features.

Acknowledgments

This work was supported by NIHR Biomedical Research Centre for Mental Health and Biomedical Research Unit for Dementia at the South London and Maudsley NHS Foundation Trust and Kings College London and a joint infrastructure grant from Guy's and St Thomas's Charity, and the Maudsley Charity; We would like to acknowledge Karl Erisman (author CLI::Framework) for his feedback on the application. We would also like to thank the authors of all the dependencies used in writing CohortExplorer and the reviewers for their feedback and suggestions.

Conflicts of Interest

None declared.

Multimedia Appendix 1

CohortExplorer source code, documentation and the Debian package.

[GZ File, 139KB - medinform_v2i2e32_app1.gz]

References

1. OpenClinica. URL: https://www.openclinica.com [accessed 2013-02-15] [WebCite Cache ID 6ESfigsjl]
2. LabKey Software. URL: http://www.labkey.com [accessed 2013-02-06] [WebCite Cache ID 6EDfuAsME]
3. Onyx. URL: http://www.obiba.org/pages/products/onyx/ [accessed 2014-11-10] [WebCite Cache ID 6TykVQ6lg]
4. Opal. URL: http://www.obiba.org/pages/products/opal/ [accessed 2014-11-10] [WebCite Cache ID 6Tykogmga]
5. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
6. Nadkarni PM, Marenco L, Chen R, Skoufos E, Shepherd G, Miller P. Organization of heterogeneous scientific data using the EAV/CR representation. J Am Med Inform Assoc 1999;6(6):478-493 [FREE Full text] [Medline: 10579606]
7. i2b2: Informatics for Integrating Biology & the Bedside. URL: https://www.i2b2.org/ [accessed 2010-08-04] [WebCite Cache ID 5rifxziRB]
8. Anhøj J. Generic design of Web-based clinical databases. J Med Internet Res 2003 Nov 4;5(4):e27 [FREE Full text] [doi: 10.2196/jmir.5.4.e27] [Medline: 14713655]
9. Nadkarni PM, Brandt C, Frawley S, Sayward FG, Einbinder R, Zelterman D, et al. Managing attribute--value clinical trials data using the ACT/DB client-server database system. J Am Med Inform Assoc 1998;5(2):139-151 [FREE Full text] [Medline: 9524347]
10. Johnson SB. Generic data modeling for clinical repositories. J Am Med Inform Assoc 1996;3(5):328-339 [FREE Full text] [Medline: 8880680]
11. Salgado NC, Gouveia-Oliveira A. Towards a common framework for clinical trials information systems. Proc AMIA Symp 2000:754-758 [FREE Full text] [Medline: 11079985]
12. Brandt CA, Morse R, Matthews K, Sun K, Deshpande AM, Gadagkar R, et al. Metadata-driven creation of data marts from an EAV-modeled clinical research database. Int J Med Inform 2002 Nov 12;65(3):225-241. [Medline: 12414020]
13. Nadkarni PM, Brandt C. Data extraction and ad hoc query of an entity-attribute-value database. J Am Med Inform Assoc 1998;5(6):511-527 [FREE Full text] [Medline: 9824799]
14. Celko J. Joe Celko's SQL for Smarties, Fourth Edition: Advanced SQL Programming (The Morgan Kaufmann Series in Data Management Systems). San Francisco (USA): Morgan Kaufmann; 2010.
15. Pennington JW, Ruth B, Italia MJ, Miller J, Wrazien S, Loutrel JG, et al. Harvest: an open platform for developing web-based biomedical data discovery and reporting applications. J Am Med Inform Assoc 2014;21(2):379-383 [FREE Full text] [doi: 10.1136/amiajnl-2013-001825] [Medline: 24131510]
16. Wade TD, Hum RC, Murphy JR. A Dimensional Bus model for integrating clinical and research data. J Am Med Inform Assoc 2011 Dec;18 Suppl 1:i96-102 [FREE Full text] [doi: 10.1136/amiajnl-2011-000339] [Medline: 21856687]
17. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.r-project.org/ [accessed 2014-11-06] [WebCite Cache ID 6TtE09kIs]
18. OBiBa: Open Source Software BioBanks. URL: http://www.obiba.org/ [accessed 2014-11-10] [WebCite Cache ID 6TyqJI4xD]
19. The Perl Programming Language. URL: https://www.perl.org/ [accessed 2014-11-10] [WebCite Cache ID 6TyqRZNzR]
20. DBI - Database independent interface for Perl. URL: https://metacpan.org/pod/DBI [accessed 2014-11-10] [WebCite Cache ID 6TyqY3n8i]
21. SQL::Abstract. URL: https://metacpan.org/pod/SQL::Abstract [accessed 2014-11-10] [WebCite Cache ID 6TyqgZ1mP]
22. CohortExplorer::Datasource. URL: https://metacpan.org/pod/CohortExplorer::Datasource [accessed 2014-11-10] [WebCite Cache ID 6TyqqNB5U]

23. CohortExplorer Demonstration (YouTube). URL: http://www.youtube.com/watch?v=Tba9An9cWDY [accessed 2014-11-07] [WebCite Cache ID 6TtEDrYEj]
24. CohortExplorer (GitHub). URL: https://github.com/abhishekdxt/CohortExplorer [accessed 2014-11-10] [WebCite Cache ID 6Tyra22hg]
25. CLI::Framework. URL: https://metacpan.org/release/CLI-Framework [accessed 2014-11-10] [WebCite Cache ID 6Tyr4BgI5]
26. Lovestone S, Francis P, Strandgaard K. Biomarkers for disease modification trials--the innovative medicines initiative and AddNeuroMed. J Nutr Health Aging 2007;11(4):359-361. [Medline: 17653500]
27. Lovestone S, Francis P, Kloszewska I, Mecocci P, Simmons A, Soininen H, AddNeuroMed Consortium. AddNeuroMed--the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. Ann N Y Acad Sci 2009 Oct;1180:36-46. [doi: 10.1111/j.1749-6632.2009.05064.x] [Medline: 19906259]
28. Kiddle SJ, Thambisetty M, Simmons A, Riddoch-Contreras J, Hye A, Westman E, Alzheimers Disease Neuroimaging Initiative. Plasma based markers of [11C] PiB-PET brain amyloid burden. PLoS One 2012;7(9):e44260 [FREE Full text] [doi: 10.1371/journal.pone.0044260] [Medline: 23028511]
29. Kiddle SJ, Sattlecker M, Proitsi P, Simmons A, Westman E, Bazenet C, et al. Candidate blood proteome markers of Alzheimer's disease onset and progression: a systematic review and replication study. J Alzheimers Dis 2014;38(3):515-531. [doi: 10.3233/JAD-130380] [Medline: 24121966]
30. BioResource. URL: http://bioresource.nihr.ac.uk/ [accessed 2014-10-24] [WebCite Cache ID 6TZMZgiLp]
31. CohortExplorer. URL: https://metacpan.org/release/CohortExplorer [accessed 2014-11-10] [WebCite Cache ID 6TyrExeBI]

Abbreviations

API: application programming interface
CDISC: Clinical Data Interchange Standards Consortium
EAV: entity-attribute-value
EDC: electronic data capture
EDM: electronic data management
i2b2: Informatics for Integrating Biology & the Bedside
NIHR: National Institute for Health Research
ODM: observational data model
SLAM: South London and Maudsley NHS Foundation Trust
SQL: structured query language

Edited by G Eysenbach; submitted 03.03.14; peer-reviewed by P Nadkarni, K Marsolo, M Italia; comments to author 02.06.14; revised version received 03.08.14; accepted 19.09.14; published 01.12.14.

Please cite as:
Dixit A, Dobson RJB
CohortExplorer: A Generic Application Programming Interface for Entity Attribute Value Database Schemas
JMIR Med Inform 2014;2(2):e32
URL: http://medinform.jmir.org/2014/2/e32/
doi:10.2196/medinform.3339
PMID:25601296

Original Paper

Adoption, Use, and Impact of E-Booking in Private Medical Practices: Mixed-Methods Evaluation of a Two-Year Showcase Project in Canada

Guy Paré[1*], PhD; Marie-Claude Trudel[1*], PhD; Pascal Forget[2*], PhD

[1]HEC Montreal, Montreal, QC, Canada

[2]Université du Québec à Trois-Rivières, Trois-Rivières, QC, Canada

[*]all authors contributed equally

Corresponding Author:
Guy Paré, PhD
HEC Montreal
3000 Chemin de la Cote-Ste-Catherine
Montreal, QC, H3T 2A7
Canada
Phone: 1 514 340 6812
Fax: 1 514 340 6132
Email: guy.pare@hec.ca

Abstract

Background: Managing appointments in private medical practices and ambulatory care settings is a complex process. Various strategies to reduce missed appointments can be implemented. E-booking systems, which allow patients to schedule and manage medical appointments online, represents such a strategy. To better support clinicians seeking to offer an e-booking service to their patients, health authorities in Canada recently invested in a showcase project involving six private medical clinics.

Objective: The objectives pursued in this study were threefold: (1) to measure adoption and use of the e-booking system in each of the clinics over a 2-year period, (2) to assess patients' perceptions regarding the characteristics and benefits of using the system, and (3) to measure the impact of the e-booking system on the number of missed appointments in each clinic.

Methods: A mixed-methods approach was adopted in this study. We first extracted and analyzed raw data from the e-booking system deployed in each of the medical practices to monitor adoption and use of the system over time and to assess the impact of the system on the number of missed appointments. Second, we conducted a Web-based survey of patients' perceptions in the spring of 2013.

Results: The patients and physicians targeted by this showcase project showed a growing interest in the e-booking system as the number of users, time slots made available by physicians, and online appointments grew steadily over time. The great majority of patients said that they appreciated the system mainly because of the benefits they derived from it, namely, scheduling flexibility, time savings, and automated reminders that prevented forgotten appointments. Importantly, our findings suggest that the system's automated reminders help significantly reduce the number of missed appointments.

Conclusions: E-booking systems seem to represent a win-win solution for patients and physicians in private medical practices. We encourage researchers to replicate and extend our work in other primary care settings in order to test the generalizability of our findings.

(JMIR Med Inform 2014;2(2):e24) doi:10.2196/medinform.3669

KEYWORDS

e-booking; medical practices; primary care; missed appointments; mixed-methods evaluative study

Introduction

One of the keys to efficiency, productivity, and profitability in private medical practices is linked to the appointment scheduling system. Managing appointments in private medical practices and ambulatory care settings is a complex process. One frequent problem faced by many clinics is related to non-attendance [1]. According to various studies, missed appointments (also called "no-shows") represent close to 10% of all medical appointments

[2,3]. There are many collateral effects associated with missed appointments for the providers, staff, and the patients themselves. For instance, no-shows can lead to lower productivity for family physicians and their staff [4]. More importantly, missed appointments increase overall wait time for all patients and can lead to additional risks to their health condition [3].

Various strategies to reduce missed appointments can be found in the extant literature [5]. One frequently mentioned approach is overscheduling, which consists of booking more appointments than the practice is actually able to accommodate [6]. While this strategy may be efficient from the standpoint of use of staff time, it usually creates a great deal of dissatisfaction for both patients and staff [7]. Another approach involves reminders, which are sent in various ways, such as by mail, telephone calls (automated or not), emails, and text messages. These are intended to minimize the risk of patients forgetting their appointments. Several studies have compared the impact of various communication methods for sending out reminders. For example, Henderson [3] observed a decrease in missed appointments when telephone or mailed reminders were used, especially when these reminders were made a few days before the appointment date. Others have observed that text message reminders are as effective as other types [8-10].

Another strategy is called advanced access scheduling [11,12]. This involves reserving appointment slots for same-day appointments, rather than booking appointment slots months in advance. In other words, physicians who use advanced access scheduling generally cut down on prescheduled visits, leaving a large portion of their day open for same-day visits. The mix between prescheduled and open appointments is usually determined by the medical practice's unique balance of supply and demand for appointments. Research has indicated that advanced access scheduling can provide numerous benefits, including increased satisfaction for patients, providers, and staff [13], fewer missed appointments [14,15], as well as increased productivity among the health care professionals [13].

E-booking systems, which allow patients to schedule and manage their medical appointments online, have also been deployed to streamline management of appointments in medical practices and ambulatory care settings [16,17]. While only 7% of Canadian family physicians (compared to 30% in the United States and 51% in Norway) offered such access in 2012 [18], 90% of surveyed Canadians in 2013 said that if the functionality were available, they would be likely to book an appointment with their health care provider electronically [19]. Survey respondents also ranked e-booking in the top three most useful online consumer health services, just behind electronic prescription renewals and viewing their lab results online. That said, when asked whether they can currently make an appointment with their family physician electronically, only 5% responded that they could.

To better support clinicians seeking to offer an e-booking service to their patients, Canada Health Infoway, a federally funded, not-for-profit organization tasked with accelerating the development of health information technologies across Canada, recently launched the *e-Booking Initiative* for eligible licensed physicians in private medical practices. This program offers financial support to help offset the costs associated with e-booking system acquisition and implementation. Canada Health Infoway also invested in a showcase project involving six private medical clinics located in Québec, Canada. The present study pursued three objectives in line with this multisite project: (1) to measure adoption (number of patients and physicians enrolled) and use (number of time slots available online, number of appointments made online) of the e-booking system in each of the clinics over a 2-year period, that is, between January 2012 and December 2013; (2) to assess patients' perceptions regarding the characteristics and benefits of the e-booking system; and (3) to measure the impact of system usage on the number of missed appointments in each participating clinic. Evidence for effective technological solutions to streamline the appointment scheduling process and improve attendance in primary care and outpatient settings is lacking. Indeed, very few empirical studies [20] have investigated the adoption, use, and effectiveness of e-booking systems in private medical practices. Hence, the present study attempts to fill this gap.

Methods

E-Booking System and Sites

The *Doctor Direct* software application (DoctorDirect.com) was deployed as part of this showcase project. This application consists of a secure Web portal that enables patients to access their doctor's schedule 24 hours a day, 7 days a week and book an appointment that suits them best without the assistance of a secretary. An email reminder, as well as a telephone reminder (automated message), are sent to the patient 2 days before the appointment. The patient is then able to confirm or cancel the appointment online. This solution was chosen because of its interoperability with the most widely used electronic medical record (EMR) system (Kinlogix Medical, TELUS Health) in medical practices in Québec [21]. The medical practices that took part in this project (see Table 1) were identified by Canada Health Infoway; they were chosen mainly because of the diversity of their profiles in terms of health care services offered and clients. Acronyms have been used to preserve anonymity of the participating clinics. It was decided that each medical practice would adopt a marketing strategy to promote the e-appointment system with its clients. As shown later, the promotion strategy for each medical practice was developed based on the patients' sociodemographic characteristics and level of comfort with the technology, as well as the preferred methods of promotion identified by the management at each site. Medical practices did not receive any financial incentives to encourage participation in this showcase project.

Table 1. Profile of the medical practices.

Medical practice	Health care services offered	Clients
A	Family medicine with two specialists on-site	Adults and children
B	Family medicine, travel health, specimen collection center, operating rooms	Adults and children
C	Transrectal echography with or without biopsy, cystoscopy, vasectomy, uroflowmetry, minor surgery, urology research	Elderly clients / primarily men
D	Medical consultation with or without an appointment, emergency and minor surgery, specimen collection service for laboratory testing, mother-child clinic, vaccination	Young families / expectant women or mothers with babies
E	General medicine	Ubisoft employees (young computer-savvy people)
F	Multidisciplinary health services	Mostly adults or elderly people

Data Collection and Analysis

A mixed-methods approach was adopted in this study. First, to monitor adoption and use of the e-booking system over the 2-year observation period, the supplier of the IT solution gave us secure access to the system's database. This allowed us to extract raw usage data from the e-booking system in use at each of the six medical practices. These data were then imported into an Excel file that was used to produce several graphs (see Results section). In line with our second objective, a Web-based questionnaire survey was conducted in the spring of 2013. Of the 4338 patients enrolled in the e-booking system at the start of the study, 1032 (23.79%) agreed to be contacted by the research team. The questionnaire, which was prepared in French and English, was posted online using *Qualtrics* software and an email invitation to take part in the study was sent to all potential respondents. A week later, an email reminder was sent to all targeted respondents. As shown in the next section, data were analyzed using various descriptive statistics (means, standard deviations) and tests (Pearson's chi-square test, Student's *t* test) as well as partial least square (PLS) multiple regression tests.

Our third and final objective was to assess the impact of the use of the e-booking system on the number of missed appointments. To this end, we began by analyzing data from Clinic A, which had recorded the most appointments made online in the period from January 1, 2012 to December 31, 2013. We compiled the number of offline appointments (made through a secretary), the number of online appointments, and the number of missed appointments (offline and online) from January 2012 to November 2013. A statistical *t* test analysis allowed us to measure the impact of the e-booking system on the number of missed appointments. Data were then collected on the four other medical clinics (B, C, D, and E) from the databases of their e-booking systems. Data for a 12-month period (December 2012 to November 2013) were analyzed, since the volume of online appointments was high enough to perform the desired analyses. Data from Clinic F were not analyzed since the volume of online appointments was too low. Data were analyzed using Student's *t* test.

Ethical approval for this study was obtained from the Research Ethics Council of HEC Montréal in March 2013.

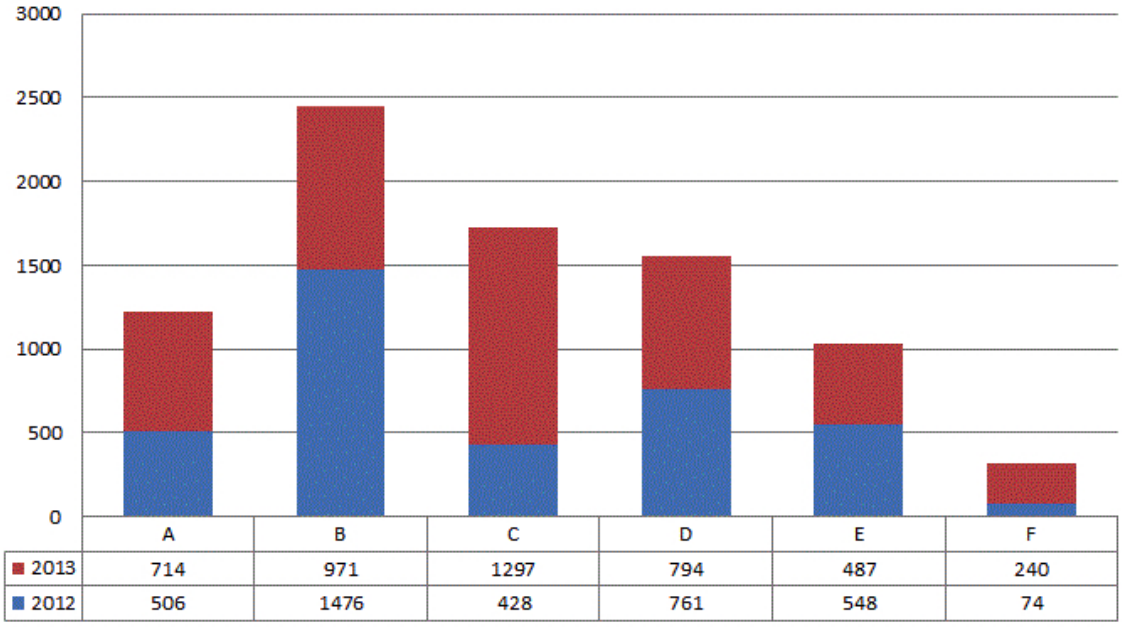
Results

Adoption and Use of the E-Booking System

The statistics presented in Figure 1 show that 8296 patients from the six medical practices enrolled with *Doctor Direct.* This represents 10.00% (3793/37,936) and 12.00% (4503/37,524) of the active patients at all six clinics in 2012 and 2013, respectively. Five of the six clinics recruited 1600 new registrants, on average, from the time they deployed the e-booking system to the end of 2013. Clinic F, which had more difficulty recruiting patients to use the system, had only 250 patients registered at the end of 2013. According to those responsible for the project, various technical problems (eg, appointment confirmations not sent, time slots offered to more than one patient), which had occurred mostly in 2012, represented an aggravating factor for this site.

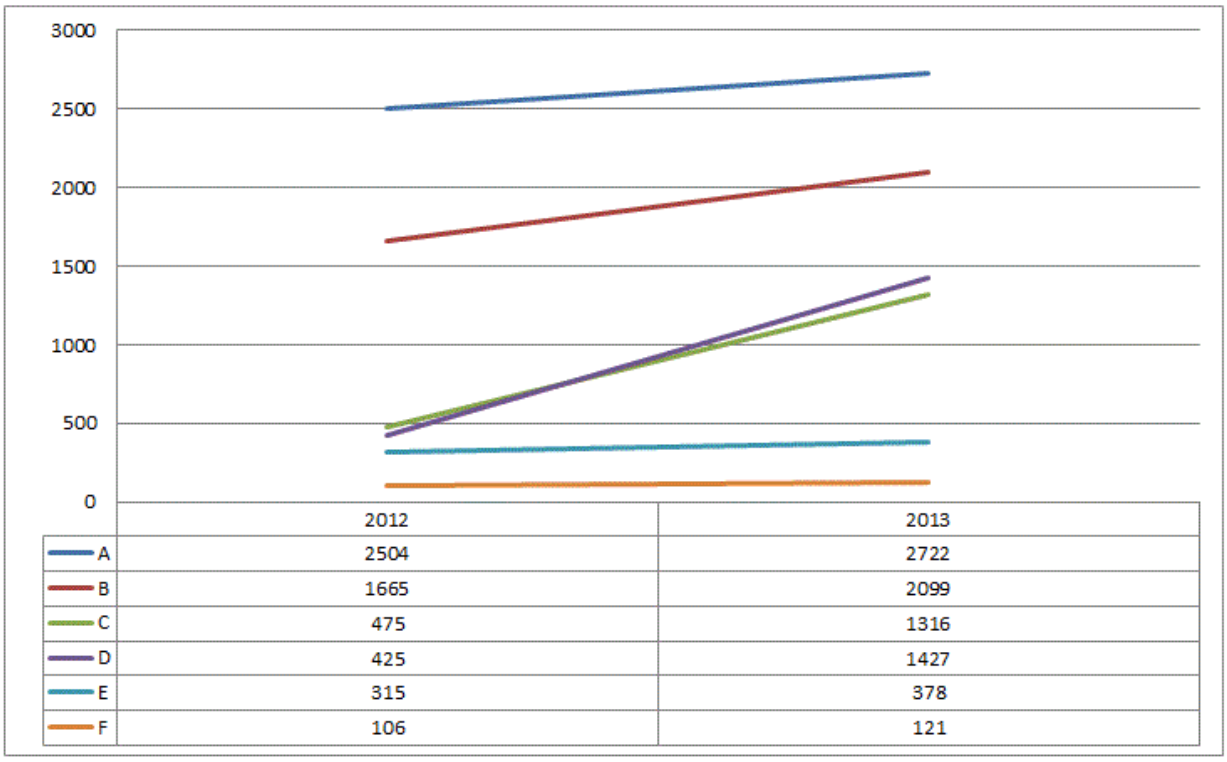
At the end of 2012, there were 34 physicians using the system in six clinics for a total of 50 possible licenses (68%). Twelve months later, 47 licenses (94%) were being used by the targeted physicians. In terms of system use, the number of time slots that the physicians had made available online grew from 23,201 in 2012 to 43,101 in 2013, for a 46% increase. As shown in Figure 2, the number of medical appointments booked online by patients grew by 32%, from 5490 in 2012 to 8063 in 2013, bringing the number of online appointments to 13,553. This represented a total of one out of every five time slots assigned to the online reservation system. Last, the average registered patient made 1.6 online appointments from the time they enrolled in the system until December 31, 2013.

Figure 1. Number of new patients enrolled, by medical practice.



	A	B	C	D	E	F
2013	714	971	1297	794	487	240
2012	506	1476	428	761	548	74

Figure 2. Number of medical appointments booked online in 2012 and 2013, by clinic.



	2012	2013
A	2504	2722
B	1665	2099
C	475	1316
D	425	1427
E	315	378
F	106	121

Survey of Patients Enrolled in the E-Booking System

A total of 228 completed questionnaires were received between March 29 and April 3. As mentioned above, a reminder letter was sent to all targeted respondents on April 4. This reminder helped retrieve an additional 147 questionnaires. The final response rate was 36.34% (375/1032), which is deemed satisfactory [22]. Among the questionnaires received, 71 had to be discarded due to missing data. The final sample was thus comprised of 304 questionnaires, including 194 received before the reminder and 110 after the reminder. As there was no statistically significant difference between early and late respondents on all attributes, response bias was unlikely [23].

As shown in Table 2, the sample consisted of two main categories of respondents: patients who had already made at least one appointment online since enrolling in the e-booking system (n=241) and patients who had not yet made an appointment using the system (n=63). The results show similarities between the two groups as to sex, age, and level of

education. The sample included slightly more women than men and all age groups were represented, although individuals aged 50 to 59 years represented the main group of respondents. Four out of five respondents had a college diploma or university degree, which shows a high level of education.

We began by asking patients who had not yet booked an appointment online (n=63) to state their reasons for not doing so. The main reason was that they had not needed to schedule a doctor's appointment between the time they enrolled and the study period (n=24). However, more than one-third of non-users (33%, 21/63) indicated that they had tried to schedule an appointment but were unable to do so because no time slot was available for their doctor. Technical problems during their first attempt discouraged only 14 respondents. It is worth mentioning that system user-friendliness and security did not seem to be major barriers to system use. We also asked this sub-group of patients the extent to which they intended to schedule their next medical appointments online. About 85% (54/63) responded positively.

We then turned our attention to patients who had booked at least one medical appointment online using *Doctor Direct* (n=241). The majority of system users (56.0%, 135/241) had booked only one appointment online, while one in four (24.0%) had booked two appointments and 20.0% had booked three or more. The vast majority (83.0%, 200/241) used the system to manage their own medical appointments, while only 17.0% (41/241) used it to book appointments for relatives. As shown in Table 3, users of the e-appointment system claimed to be very satisfied (average of 4.2 on a scale of 5), perceived the system as very user-friendly (4.3/5), and had a firm intention of continuing to use it in the future (4.5/5).

To further investigate the factors that motivate patients to continue using the e-appointment system in the future, we tested a research model derived from the works of Bhattacherjee [24] and Hong et al [25] on information systems continuance. As shown in Figure 3, our model suggests that an individual's intention to continue using a computer-based system is mainly influenced by his or her level of satisfaction toward the system. In turn, user satisfaction is influenced by the extent to which initial expectations toward the system are confirmed as well as by two factors from the TAM (technology acceptance model) proposed by Davis [26], namely, system ease of use and system usefulness. Following Hong et al [25], our model also proposes direct links between the TAM constructs and the dependent variable. The survey instrument that was used is presented in Multimedia Appendix 1. The reliability of the measures was determined with Cronbach alpha. Findings in Table 3 indicate that all the measures, without exception, meet or surpass the .70 threshold of statistical significance [27]. This table also demonstrates the validity of the variables included in our research model. In particular, we see that the square root of the variance shared by each variable and its respective items is greater than the inter-correlations between the variables.

PLS regression analyses were performed to test the links in our model. Our findings supported all relationships, with the exception of the association between system ease of use and continuance intention. It would thus appear that system user-friendliness has an indirect effect on the dependent variable via its direct influence on user satisfaction. Most importantly, our findings underline the importance of the "expectation confirmation" variable which, as anticipated, is strongly related to TAM factors and user satisfaction. This result shows the importance of managing users' initial expectations to ensure that they are not disappointed when they first attempt to use the system.

Next, Table 4 indicates that three kinds of benefits were perceived by system users: scheduling flexibility, time savings, and automated reminders that prevented forgotten appointments.

Concerning the marketing or promotional strategies implemented in each medical clinic, we asked all respondents (n=304) to indicate what had led them to enroll in the e-booking system. As shown in Table 5, half of them mentioned that they enrolled because a secretary had recommended it during a prior visit to the clinic. One out of five patients signed on to the Internet portal at the recommendation of their physician, and approximately 15% were inspired by the message on the clinic's voicemail and the tab on the medical clinic's website. The brochures and posters promoting the portal in the clinics' waiting rooms appeared to have had little effect on enrollments, since they were mentioned by only 6% of respondents. No significant statistical differences were found across medical practices.

Table 2. Profile of survey respondents (n=304).

		Patients who booked online at least once (n=241) n (%)	Patients yet to book online (n=63) n (%)	χ^2 and *t*	*P* value
Sex					
	Men	109 (45.2)	22 (34.9)		
	Women	131 (55.4)	39 (61.9)	χ^2=1.7	.197
Age, years					
	18–29	21 (8.7)	8 (12.7)		
	30–39	60 (24.9)	13 (20.6)		
	40–49	29 (12.0)	8 (12.7)		
	50–59	66 (27.4)	18 (28.6)		
	60–69	46 (19.1)	11 (17.5)		
	70+	19 (7.8)	5 (7.9)	χ^2=1.3	.933
Education					
	None	4 (1.7)	0 (0.0)		
	High school diploma	44 (18.3)	10 (15.9)		
	College diploma	54 (22.4)	16 (25.4)		
	Bachelor degree	73 (30.2)	26 (41.3)		
	Master’s degree	53 (22.0)	7 (11.1)		
	PhD	12 (5.0)	3 (4.8)	χ^2=6.3	.279
Medical practices					
	A	18 (7.5)	6 (9.5)		
	B	70 (29.0)	39 (61.9)		
	C	77 (32.0)	2 (3.2)		
	D	57 (23.7)	7 (11.1)		
	E	13 (5.4)	6 (9.5)		
	F	6 (2.5)	3 (4.8)	χ^2=55.1	.000
Level of computer knowledge [a]		4.5	4.2	*t*=2.3	.022

[a]Scale of 1 to 5 where 1=slightly familiar and 5=very familiar.

Table 3. Descriptive statistics and variance shared by the variables.

	Mean	SD	Number of items	Cronbach alpha	PU	EOU	CONF	SAT	CONT
Perceived usefulness of the system (PU)	4.2	0.9	4	.86	.85 [a]				
User-friendliness of the system (EOU)	4.3	0.8	4	.93	.68[b]	.91			
Confirmation of expectations (CONF)	4.0	1.0	3	.87	.82[b]	.68[b]	.89		
Satisfaction with the system (SAT)	4.2	0.9	4	.80	.72[b]	.58[b]	.73[b]	.82	
Intention to continue using the system (CONT)	4.5	0.8	3	.93	.81[b]	.62[b]	.76[b]	.72[b]	.94

[a]The ratios on the diagonal represent the square root of the variance shared by each variable and its respective items. The ratios below the diagonal are correlations between variables.

[b]*P*<.001.

Table 4. Perceived benefits of using the e-booking system (n=241).

		Average (1-5 scale)	SD
Greater flexibility			
	Makes it possible to book appointments when it is most convenient.	4.7	1.0
	Greater flexibility in the choice of available time slots.	4.6	1.2
Time savings			
	Saves time by eliminating waiting on the phone.	4.5	1.0
	Saves time by eliminating the need for reminders several times at the clinic when the phone is busy.	4.5	1.2
	Saves time by eliminating the need for me to go in person to the clinic to schedule an appointment.	4.5	1.2
Reduction in forgotten appointments			
	Makes it easier to remember appointments thanks to reminders.	4.5	1.0

Table 5. Promotional strategies put in place and patients' receptiveness (n=304).

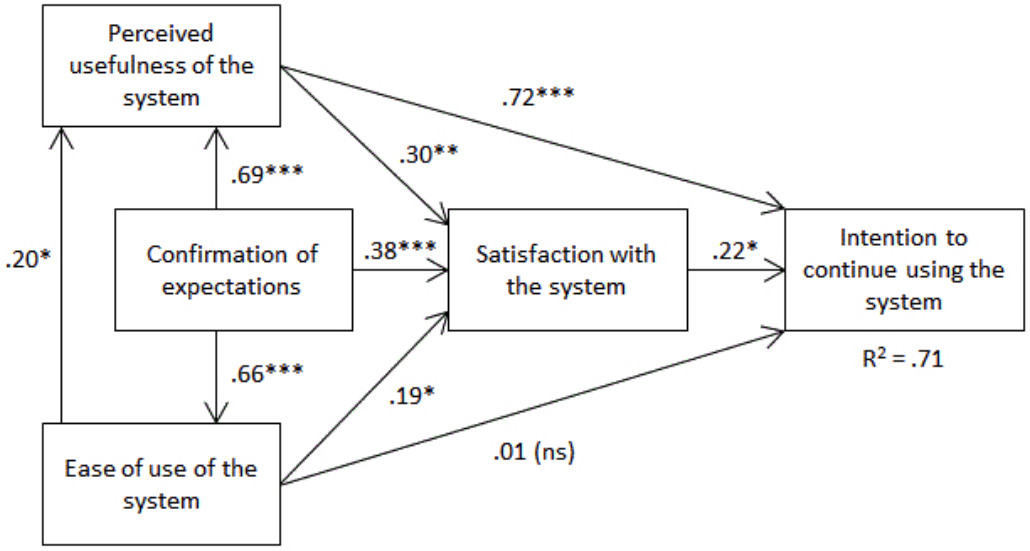
Promotional strategy	Clinic A	Clinic B	Clinic C	Clinic D	Clinic E	Clinic F	Patients who were influenced, n (%)
Secretary's verbal recommendation	√[a]	√	√	√	√	√	158 (52.0)
Physician's verbal recommendation	√	√	√	√	√	√	62 (20.3)
Promotional message on the clinic's voicemail	√	√	√	√	√	√	49 (16.1)
Link on the medical clinic's website		√	√	√		X	45 (14.8)
Flyer distributed at the medical clinic	√	√	√	√	√	√	21 (6.9)
Promotional poster in the medical clinic	√	√	√	√	√	√	17 (5.6)
Interactive terminals available in the clinic (iPads)	X[b]		X	√		X	-
Email invitation to all patients					√	√	N/A[c]

[a]√ = Strategy implemented before the survey conducted in the spring of 2013.

[b]X=Strategy implemented after the survey conducted in the spring of 2013.

[c]N/A=Data not available in the survey questionnaire.

Figure 3. Research model and PLS results (n=241). ***P<.005; **P<.01; *P<.05; ns=not significant.

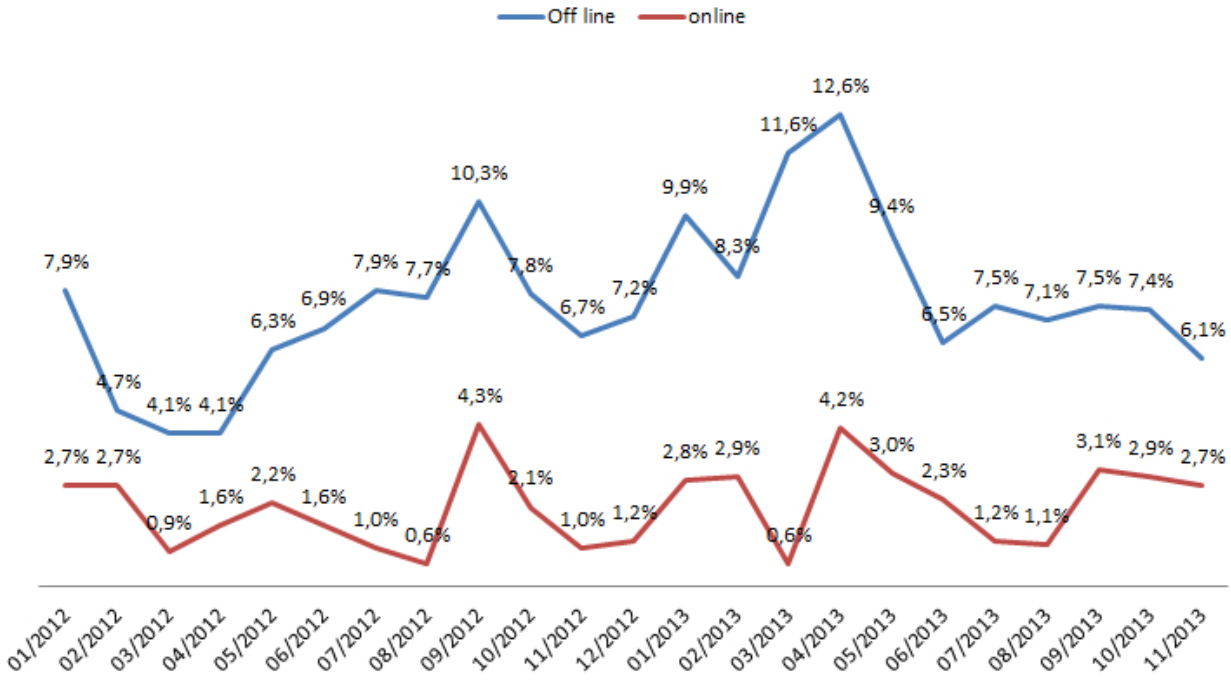


Impact of E-Booking on the Number of Missed Appointments

Our third and final goal was to assess the impact of the e-booking system on the number of missed appointments. As explained above, we began by analyzing Clinic A's data. The results shown in Figure 4 indicate that the percentage of missed appointments each month varied from 3.4% to 11% and averaged 6.5%. However, when we compare appointments made online (by the patients themselves) from those made offline, we note a large difference in the number of missed appointments. The percentage of online appointments that were missed varied from 0.6% to 4.3%, averaging 2.1%. Considering appointments made in the traditional manner, missed appointments represented 4.1% to 12.6% of the total and averaged 7.6%. The difference between the two groups (offline and online) in terms of the number of missed appointments is statistically significant (t=8.8; P<.001).

Similar results were then obtained from four other medical practices over a 12-month observation period from December 2012 to November 2013: Clinic B (t=6.3; P<.001), Clinic C (t=5.8; P<.001), Clinic D (t=4.0; P<.005), and Clinic E (t=2.2; P<.05).

Figure 4. Proportion of missed appointments at medical practice A.



Discussion

Principal Findings

Overall, the patients targeted by this showcase project showed a growing interest in the e-booking system as the number of users grew steadily over time. The promotion strategies that had greatest impact on the number of enrollments were verbal recommendations from a secretary and, to a lesser extent, from the attending physician. The great majority of users said that they appreciated the system because they found it user-friendly and for the benefits they derived from it, and this can be seen in the constantly increasing number of appointments made online over the 2-year period. Three main categories of benefits were perceived by patients, namely, scheduling flexibility, time savings, and automated reminders that prevented forgotten appointments. Findings also reveal that the number of time slots opened up by the physicians also grew month after month, and this represents a critical success factor [16]. Indeed, those respondents who had tried to schedule an appointment but were unable to do so because no time slot was available for their doctor are among those who had no intention of continuing to use the e-booking in the future. Last, in line with prior findings [20], our study reveals that the use of an e-booking system can help significantly reduce the number of no-shows or missed appointments.

Despite the encouraging results presented above, some physicians were still hesitant to make time slots available online. One reason cited by our respondents was related to the fact that there are different types of medical appointments (eg, routine annual examinations, prenatal check-ups, surgical follow-up), and they vary in length. This constraint was discussed during the project, and a strategy was developed in response: the development of pop-up menus. Such menus act as filters that, through structured questions (eg, the patient's first appointment: yes/no, a diagnosis requiring follow-up, etc.), lead the patient to select the right type of appointment, that is, one for the right amount of time. In addition to this solution, we believe that better integration of the e-booking system into the EMR system used by each clinic could facilitate the allocation of time slots by adapting the type of time slot to the health condition of each patient. Last, it is important to manage physicians' expectations. If a physician has not freed up a sufficient number of time slots for online appointments, patients may lose interest and stop using the system. Setting realistic objectives by carefully targeting the percentage of time slots to be offered online and/or by beginning with specific types of appointments (eg, vaccination clinics or short, regular follow-up appointments) may encourage a gradual transition to routine system use.

With regard to promotional strategies, secretaries and physicians must continue to encourage patients to use the e-booking system, particularly since such use leads to a significant decline in missed appointments. It would appear important to emphasize the benefits of system use: flexibility in making appointments, the time saved, and automated reminders, which prevent patients from forgetting their appointments, rather than the system's features, such as its user-friendliness, security, and reliability. Another suggestion would be to send periodic reminders to patients enrolled in the system so that they will not forget about the system and about having enrolled in it. These reminders should clearly present how to recover forgotten user codes and passwords. To prevent these messages from being perceived as junk mail and ignored, they could be combined with general information designed to make patients more responsible for their health or by public health messages.

Limitations

The results of this study must be interpreted with caution due to its inherent limitations. For one thing, we are mindful of the small scale of the showcase project. Future studies should try to validate our findings among a larger number of medical practices and contexts. We also recognize the usual constraints and generalization limitations associated with cross-sectional surveys [22]. Next, it is important to mention, with respect to generalization, that our survey was limited, as we were unable to estimate the characteristics of the reference population. This is a direct consequence of using, as a recruitment strategy, voluntary participation for completing an online questionnaire. Importantly, we analyzed secondary and survey data associated with a single e-booking system that necessarily has its own characteristics. Our findings must therefore be replicated with other e-booking platforms. Last, it would also be interesting to carry out in-depth interviews with actual users (both patients and physicians) of e-booking systems so as to gain a richer insight into the data obtained through survey questionnaires.

Conclusions

In short, the main purpose of this study was to assess perceived and actual outcomes following the deployment of an e-booking system in six medical practices in Canada. Our results show that e-booking systems seem to represent a win-win solution for patients and physicians. For one thing, patients appreciate using such a system due to its flexibility and the fact that use allows them to save time. Further, our analyses suggest that the system's automated reminders help significantly reduce the number of missed appointments, a problem that plagues several medical practices. We encourage health informatics researchers to replicate and extend our work in other primary care settings in order to test the generalizability of our results.

Acknowledgments

We would like to thank all the patients who took part in the questionnaire survey. Our thanks as well go to Louis-Charles Grano for his contribution to designing the questionnaire and collecting and analyzing the data. We would also like to thank Josée Maringo, Abhishek Kumar, and Yuanchun Song, who developed the application that enabled patients to enroll in the e-booking system and indicate whether or not they wished to be contacted by the research team, as well as the Web interface that gave us access to the raw data in the *Doctor Direct* databases. We would also like to acknowledge the contribution made by Alexandre

Ducharme and Thomas Micheneau during the data extraction process. We extend warm thanks to Karine Blondin and Sabrina Boutin for their rapid responses to our many requests, especially those related to the promotional strategies used by the various clinics and for the support that they gave us throughout this research project. We would like to express a big thank you as well to Nichad Datoo and Éric Bourbeau for their invaluable advice and assistance, especially for having provided access to weekly follow-up reports on the project and data on missed appointments. Last, we would also like to acknowledge the financial support of Canada Health Infoway.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Survey instrument.

[PDF File (Adobe PDF File), 71KB - medinform_v2i2e24_app1.pdf]

References

1. George A, Rubin G. Non-attendance in general practice: a systematic review and its implications for access to primary health care. Fam Pract 2003 Apr;20(2):178-184 [FREE Full text] [Medline: 12651793]
2. Lowes R. How to handle no-shows: getting tough is not enough. You've got to discover—and try to eliminate—the reasons why patients skip appointments. Medical Economics 2005;82(3) [FREE Full text]
3. Henderson R. Encouraging attendance at outpatient appointments: can we do more? Scott Med J 2008 Feb;53(1):9-12. [Medline: 18422203]
4. Boyette B, Staley-Sirois M. Divurgent. 2012 Mar 01. Clinical no-show rates: Is technology a contributor? URL: http://divurgent.com/volume-2-edition-2-clinical-no-show-rates-is-technology-a-contributor/ [accessed 2014-09-16] [WebCite Cache ID 6Sdb74GNb]
5. Johnson BJ, Mold JW, Pontious JM. Reduction and management of no-shows by family medicine residency practice exemplars. Ann Fam Med 2007;5(6):534-539 [FREE Full text] [doi: 10.1370/afm.752] [Medline: 18025491]
6. King A, David D, Jones HS, O'Brien C. Factors affecting non-attendance in an ophthalmic outpatient department. J R Soc Med 1995 Feb;88(2):88-90 [FREE Full text] [Medline: 7769601]
7. Torgerson D. Non-attendance in outpatients. J R Soc Med 1995 Jun;88(6):364 [FREE Full text] [Medline: 7629780]
8. Hasvold PE, Wootton R. Use of telephone and SMS reminders to improve attendance at hospital appointments: a systematic review. J Telemed Telecare 2011;17(7):358-364 [FREE Full text] [doi: 10.1258/jtt.2011.110707] [Medline: 21933898]
9. Taylor NF, Bottrell J, Lawler K, Benjamin D. Mobile telephone short message service reminders can reduce nonattendance in physical therapy outpatient clinics: a randomized controlled trial. Arch Phys Med Rehabil 2012 Jan;93(1):21-26. [doi: 10.1016/j.apmr.2011.08.007] [Medline: 22000821]
10. Sims H, Sanghara H, Hayes D, Wandiembe S, Finch M, Jakobsen H, et al. Text message reminders of appointments: a pilot intervention at four community mental health clinics in London. Psychiatr Serv 2012 Feb 1;63(2):161-168. [doi: 10.1176/appi.ps.201100211] [Medline: 22302334]
11. Anderson JB, Sotolongo CA. Implementing advanced access in a family medicine practice: a new paradigm in primary care. N C Med J 2005;66(3):223-225. [Medline: 16130949]
12. Qu X. Development of appointment scheduling rules for open access scheduling. Indiana, United States: Purdue University; 2006. URL: http://docs.lib.purdue.edu/dissertations/AAI3259977/ [accessed 2014-09-17] [WebCite Cache ID 6SerFRNhF]
13. Belardi FG, Weir S, Craig FW. A controlled trial of an advanced access appointment system in a residency family medicine center. Fam Med 2004 May;36(5):341-345 [FREE Full text] [Medline: 15129381]
14. Cameron S, Sadler L, Lawson B. Adoption of open-access scheduling in an academic family practice. Can Fam Physician 2010 Sep;56(9):906-911 [FREE Full text] [Medline: 20841595]
15. Rose KD, Ross JS, Horwitz LI. Advanced access scheduling outcomes: a systematic review. Arch Intern Med 2011 Jul 11;171(13):1150-1159 [FREE Full text] [doi: 10.1001/archinternmed.2011.168] [Medline: 21518935]
16. Green J, McDowall Z, Potts HW. Does Choose & Book fail to deliver the expected choice to patients? A survey of patients' experience of outpatient appointment booking. BMC Med Inform Decis Mak 2008;8:36 [FREE Full text] [doi: 10.1186/1472-6947-8-36] [Medline: 18673533]
17. Azouzi R, Forget P, D'Amours S. Framework for e-appointment systems design. 2012 Presented at: Proceedings of the 4th International Conference on Information Systems, Logistics and Supply Chain; 2012; Quebec City, Quebec, Canada p. 1-8.
18. Schoen C, Osborn R, Squires D, Doty M, Rasmussen P, Pierson R, et al. A survey of primary care doctors in ten countries shows progress in use of health information technology, less in other areas. Health Aff (Millwood) 2012 Dec;31(12):2805-2816. [doi: 10.1377/hlthaff.2012.0884] [Medline: 23154997]

19. Canada Health Infoway. Consumer health solutions: exploring the value, benefits and common concerns of e-booking. 2014 Mar. URL: https://www.infoway-inforoute.ca/index.php/programs-services/investment-programs/consumer-health-solutions/e-booking-initiative [accessed 2014-09-16] [WebCite Cache ID 6SdbmRAsJ]
20. Horvath M, Levy J, L'Engle P, Carlson B, Ahmad A, Ferranti J. Impact of health portal enrollment with email reminders on adherence to clinic appointments: a pilot study. J Med Internet Res 2011;13(2):e41 [FREE Full text] [doi: 10.2196/jmir.1702] [Medline: 21616784]
21. Paré G, Ortiz de Guinea A, Raymond L, Poba-Nzaou P, Trudel MC, Marsan J, et al. Canada Health Infoway Report. 2013. Computerization of primary care medical clinics in Quebec: Results from a survey on EMR adoption, use and impacts URL: https://www.infoway-inforoute.ca/ [accessed 2014-09-16] [WebCite Cache ID 6Sdc2mVKP]
22. Pinsonneault A, Kraemer KL. Survey research methodology in management information systems: an assessment. Journal of Management Information Systems 1993;10(2):75-105.
23. Scott A, Jeon SH, Joyce CM, Humphreys JS, Kalb G, Witt J, et al. A randomised trial and economic evaluation of the effect of response mode on response rate, response bias, and item non-response in a survey of doctors. BMC Med Res Methodol 2011;11:126 [FREE Full text] [doi: 10.1186/1471-2288-11-126] [Medline: 21888678]
24. Bhattacherjee A. Understanding information systems continuance: an expectation-confirmation model. MIS Quarterly 2001;25(3):351-370.
25. Hong S, Thong JYL, Tam KY. Understanding continued information technology usage behavior: A comparison of three models in the context of mobile internet. Decision Support Systems 2006 Dec;42(3):1819-1834. [doi: 10.1016/j.dss.2006.03.009]
26. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Management Science 1989;35(8):982-1003.
27. Nunnally J. Psychometric Methods. New York: NY: McGraw-Hill; 1978.

Abbreviations

EMR: electronic medical record
PLS: partial least square
TAM: technology acceptance model

Edited by G Eysenbach; submitted 03.07.14; peer-reviewed by A Shachak, H Yang, KM Augestad; comments to author 28.07.14; revised version received 20.08.14; accepted 09.09.14; published 24.09.14.

Please cite as:
Paré G, Trudel MC, Forget P
Adoption, Use, and Impact of E-Booking in Private Medical Practices: Mixed-Methods Evaluation of a Two-Year Showcase Project in Canada
JMIR Med Inform 2014;2(2):e24
URL: http://medinform.jmir.org/2014/2/e24/
doi:10.2196/medinform.3669
PMID:25600414

Viewpoint

Making Big Data Useful for Health Care: A Summary of the Inaugural MIT Critical Data Conference

Omar Badawi[1*], PharmD, MPH; Thomas Brennan[1*], PhD; Leo Anthony Celi[1*], MD, MPH, MS; Mengling Feng[1*], PhD; Marzyeh Ghassemi[1*], MS; Andrea Ippolito[1*], MS, MEng; Alistair Johnson[1*]; Roger G Mark[1*], MD, PhD; Louis Mayaud[1*], PhD; George Moody[1*]; Christopher Moses[1*]; Tristan Naumann[1*], MS; Vipan Nikore[1*], MD, MBA; Marco Pimentel[1*], MS; Tom J Pollard[1*]; Mauro Santos[1*]; David J Stone[1*], MD; Andrew Zimolzak[1*], MD, MS; MIT Critical Data Conference 2014 Organizing Committee[1*]

MIT Critical Data Conference 2014 Organizing Committee, Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, MA, United States

[*]all authors contributed equally

Corresponding Author:
Leo Anthony Celi, MD, MPH, MS
MIT Critical Data Conference 2014 Organizing Committee
Institute for Medical Engineering & Science
Massachusetts Institute of Technology
77 Massachusetts Avenue
E25-505
Cambridge, MA, 02139
United States
Phone: 1 617 253 7937
Fax: 1 617 258 7859
Email: lceli@mit.edu

Related Article:

This is a corrected version. See correction statement: http://medinform.jmir.org/2015/1/e6/

Abstract

With growing concerns that big data will only augment the problem of unreliable research, the Laboratory of Computational Physiology at the Massachusetts Institute of Technology organized the Critical Data Conference in January 2014. Thought leaders from academia, government, and industry across disciplines—including clinical medicine, computer science, public health, informatics, biomedical research, health technology, statistics, and epidemiology—gathered and discussed the pitfalls and challenges of big data in health care. The key message from the conference is that the value of large amounts of data hinges on the ability of researchers to share data, methodologies, and findings in an open setting. If empirical value is to be from the analysis of retrospective data, groups must continuously work together on similar problems to create more effective peer review. This will lead to improvement in methodology and quality, with each iteration of analysis resulting in more reliability.

(JMIR Med Inform 2014;2(2):e22) doi:10.2196/medinform.3447

KEYWORDS

big data; open data; unreliable research; machine learning; knowledge creation

Introduction

Failure to store, analyze, and utilize the vast amount of data generated during clinical care has restricted both quality of care and advances in the practice of medicine. Other industries, such as finance and energy, have already embraced data analytics for the purpose of learning. While such innovations remain relatively limited in the clinical domain, interest in “big data in clinical care” has dramatically increased. This is due partly to the widespread adoption of electronic medical record (EMR) systems and partly to the growing awareness that better data analytics are required to manage the complex enterprise of the health care system. For the most part, however, the clinical

enterprise has not had to address the problems particular to "big data" because it has not yet satisfactorily addressed more fundamental data management issues. It is now becoming apparent that we are on the cusp of a great transformation that will incorporate data and data science integrally within the health care domain. In addition to the necessary major digital enhancements of the retrospective analyses that have variably been in place, real time and predictive analytics will also become ubiquitous core functionalities in the more firmly data-based environment of the (near) future. The initial Massachusetts Institute of Technology (MIT) Critical Data Conference was conceived and conducted to address the many data issues involved in this important transformation [1,2].

Increasing interest in creating the clinical analog of "business intelligence" has made evident the necessity of developing and nurturing a clinical culture that can manage and translate data-based findings, including those from "big data" studies. Combining this improved secondary use of clinical data with a data-driven approach to learning will enable this new culture to close the clinical data feedback loop facilitating better and more personalized care. Authors have noted several hallmarks of "big data": very large datasets, a large number of unrelated and/or unstructured datasets, or high speed or low latency of data creation [3,4]. The intensive care unit (ICU) provides a potent example of a particularly data rich clinical domain with the potential for both clinical and financial benefits if these large amounts of data can be harnessed and systematically leveraged into guiding practice. Thus, we use the term "Critical Data" to refer to big data in the setting of the ICU.

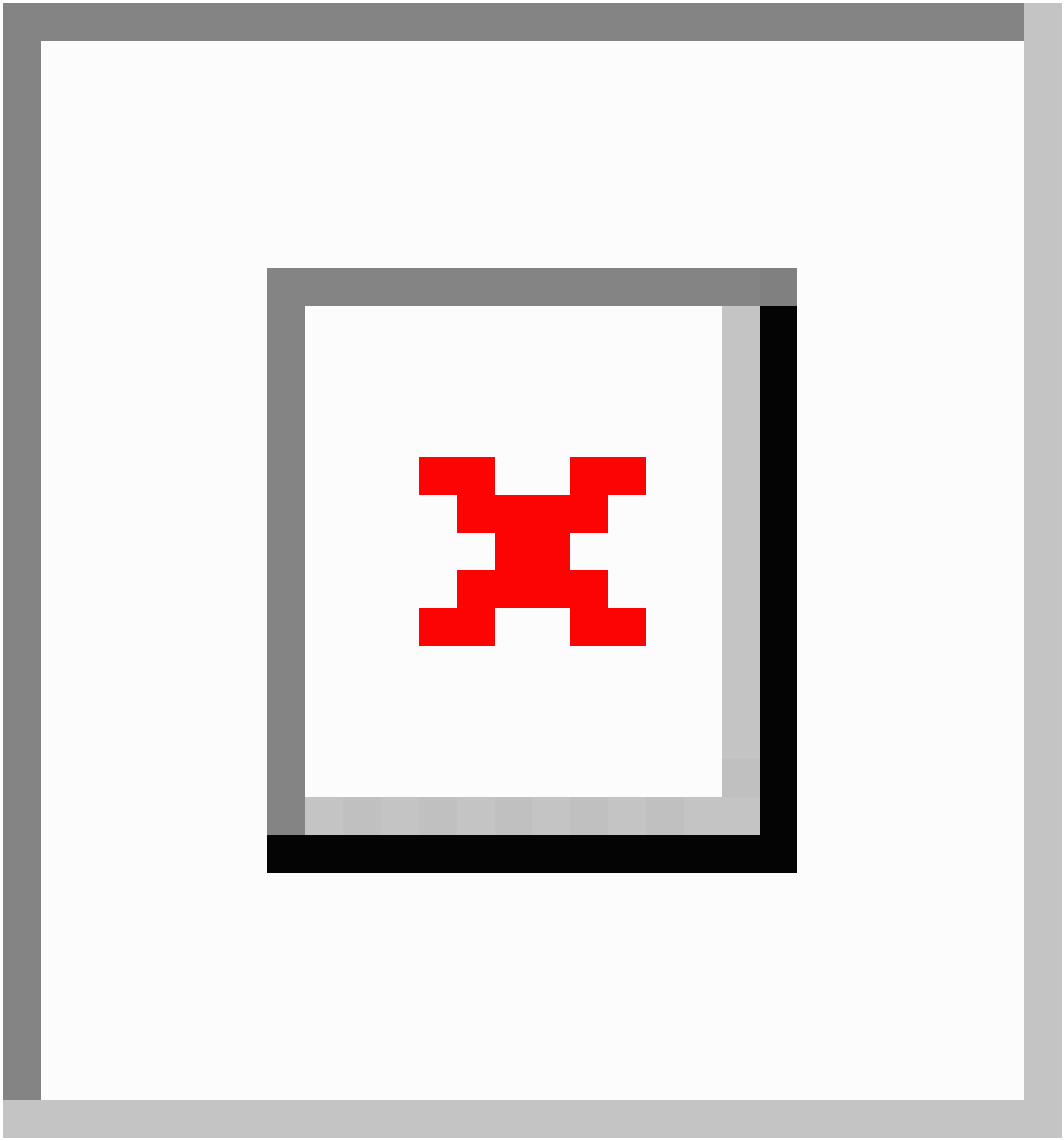
This paper summarizes the lectures and group discussions that took place during the recent Critical Data Conference at MIT, Cambridge MA, on January 7, 2014. The conference was the second part of a two-part event that brought together clinicians, data scientists, statisticians and epidemiologists.

The event opened with a "data marathon" on January 3-5, 2014 (Figure 1), which brought together teams of data scientists and clinicians to mine the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) database (version II). MIMIC II is an open-access database consisting of over 60,000 recorded ICU stays from the adult intensive care units at the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA [5]. Over 100 people participated in the two-day data marathon, and posters of the projects were displayed at the Critical Data Conference.

The Critical Data Conference on January 7 was an approximately ten-hour program comprising two keynote addresses (Jeffrey Drazen, MD and John Ioannidis, MD, PhD), seven individual lectures, three panel discussions, and two poster sessions (Figure 2). The overall conference theme was meaningful secondary use of big data from critical care settings. Materials from the conference (program, slides, and videos) are available online at the MIT Critical Data conference site [6].

Figure 1. Presentation at the Critical Data Marathon. Photo credit: Andrew Zimolzak.

Figure 2. Critical Data poster session. Photo credit: Andrew Zimolzak.



The Problem

In his keynote address, Jeffrey Drazen, MD, Editor-in-Chief of the New England Journal of Medicine, noted that the number of evidence-based recommendations built on randomized controlled trials (RCTs), the current gold standard for data quality, is insufficient to address the majority of clinical decisions. Subsequently, clinicians are often left to practice medicine "blindly." Without the knowledge generation required to capture the decisional factors involved in realistic clinical scenarios, clinical decision making is often less data-driven than determined by the "play of chance" buttressed by past experience. Historically, a doctor took a history, performed a physical examination and made a diagnosis based on what he or she observed. As technology and medical theory progressed, knowledge such as laboratory and imaging modalities helped mitigate chance in the diagnosis of disease. Rote application of existing knowledge is not enough, as physicians want to establish causality. Until now this has been done with theories, but moving forward, theories will be inadequate unless they are confirmed, translated to practice, and systematically disseminated in clinical practice.

This trial-and-error process continues today because data generated from routine care is most often not captured and is rarely disseminated for the purpose of improving population health. Even in information-rich care settings like the ICU, the knowledge necessary to mitigate the play of chance is lacking [7,8]. As such, the ICU provides a fertile ground for potential

improvement. Specifically, Drazen suggested a potential role for clinical data mining to answer questions that cannot be answered using RCTs [9]. This approach would likely yield benefits both more quickly and with fewer resources.

Drazen concluded with the question "At what point is data good enough?" Documented associations may be strong but not sufficiently "proven" to establish causality. Drazen drew a comparison with experimental physicists who hone future studies on the work of theorists as well as on prior experimental results: biomedical informaticians can identify meaningful associations that can then guide design of new RCTs where data quality can be increased by controlling for potential confounders. This will require cross-disciplinary collaboration of frontline clinicians, medical staff, database engineers and biomedical informaticians, in addition to strong partnerships with health information system vendors in order to close the loop from knowledge discovery during routine care to the real time application of best care for populations.

Secondary Usage of Clinical Data

Charles Safran, MD, MS, Chief of the Division of Clinical Computing at the BIDMC and Harvard Medical School, spoke next, sharing the dream of evidence-based medicine (EBM): the ideal situation in which quality evidence would exist to guide clinicians through all the conundrums faced on a near-daily basis (eg, which test to order, how to interpret the test results, and what therapy to institute). For the last half-century, prospective RCTs have been the gold standard in EBM. Safran noted, as Drazen had, that such trials suffer from a number of limitations including economic burdens and design limitations. An RCT can only address a severely limited bundle of particularly well-posed clinical questions. For many clinical situations, it is either unethical or even impossible to proceed with an RCT. Furthermore, the inclusion and exclusion criteria often limit the generalizability of an RCT study and, given the time it normally takes to run an RCT, it is very difficult for these studies to remain current with the rapidly evolving practice of medicine.

Can another approach avoid at least some of the limitations of RCTs? Safran suggested that retrospective observational studies (ROS) utilizing EMR data are a promising avenue for generating EBM. Digital records contain extensive clinical information including medical history, diagnoses, medications, immunization dates, allergies, radiology images, and laboratory and test results. Consequently, routinely collected EMR data contains the rich, continuous and time-sensitive information needed to support clinical decision making and evidence generation [10]. However, despite the many potential benefits, Safran pointed out that secondary use of EMR data is still subject to limitations: EMR data were not collected primarily for the purpose of evidence generation and data analytics but for real-time and longitudinal patient care [11]. As a result, EMR data are often poorly structured, disorganized, unstandardized, and contaminated with errors, artifacts, and missing values.

Safran echoed one of Drazen's points by proposing that we should combine the usage of prospective RCTs and ROS in such a way that each complements the limitations of the other. Furthermore, he suggested the possibility of incorporating additional novel sources of data such as social media data, health data from portable sensors, and genetic data. While there are many barriers to establishing such a comprehensive framework, a big data picture of clinical, genetic, and treatment variables holds promise in revolutionizing diagnosis and treatment.

Connecting Patients, Providers, and Payers

For John Halamka, MD, MS, the Chief Information Officer of BIDMC, working with big data in hospital systems is hugely challenging but at the same time holds tremendous promise in providing more meaningful information to help clinicians treat patients across the continuum of care. In his position, Halamka has been tasked to aggregate data in novel ways in order to provide better care for BIDMC's patient population. One opportunity for furthering "big data in health care" is to normalize the data collected via their EMR system and store it in large, centralized databases. In turn, analytic tools can then be applied to identify and isolate the quality data reporting measures required to participate as an Accountable Care Organization (ACO) under the Affordable Care Act.

Halamka emphasized that building these large datasets does not intrinsically provide value from the start, stating that "workflow is disparate, the vocabulary is disparate, and the people are disparate." Therefore, the normalization of data and its distillation into standard schemas are difficult due to discrepancies across longitudinal data. Further, since each vendor models concepts differently, there must be an emphasis on developing a "least common denominator" concept map across vendors' offerings.

Nevertheless, through this normalization effort, doctors can utilize "scorecards" to evaluate their own patient population within and across the different payment models, such as Blue Cross Blue Shield's Alternative Quality Contact measures, the Center for Medicare and Medicaid (CMS) Physician Quality Reporting System measures, and the CMS ACO measures. In addition, physicians can query this dataset to identify the most effective treatment regimes. However, such queries do pose privacy and security issues in the hospital setting, and these risks are further complicated by hospital staff utilizing personal mobile devices such as cell phones, laptops, and tablets.

Creating a Data-Driven Learning System

The problem posed to the first panel (Figure 3), comprising Gari Clifford, PhD, Perren Cobb, MD, and Joseph Frassica, MD, and moderated by Leo Anthony Celi, MD, MS, MPH, was how to create a data-driven learning system in clinical practice [8]. Privacy concerns were cited as the central barrier, as there is a tradeoff between re-identification risk and the value of sharing. Furthermore, recent work shows patients are reluctant to share for certain purposes such as marketing, pharmaceutical, and quality improvement measures, indicating a need for public education about the benefits of data sharing and that shared data can be utilized without being used for marketing and other unwanted purposes [2].

There is also tension between intellectual property rights and transparency. Resolution of this may require collaboration between government, industry and academic institutions, as seen with the US Critical Illness and Injury Trials Group [12]. There is also a risk that data sharing will make authors reluctant to write audacious or unconventional papers (as did Reinhart and Rogoff [13]), if data sharing puts such papers at perceived higher risk of refutation (such as the refutation of Herndon et al [14]).

Finally, the panel raised concerns about the high quantity but perceived low quality of the data that is actually captured. While there is hope that automatically captured data may be more accurate than manually entered data, there is also some risk that doing so will introduce additional noise, furthering the problem of quantity over quality. This concern poses the challenge of capturing more and higher quality data in order to promote reproducibility. Panelists observed that multidisciplinary conferences like the Critical Data Conference are especially beneficial in this regard, as they provide an opportunity for clinicians and data scientists to better understand the relation between real-world activity and the data that such activity generates.

Figure 3. Data-driven learning system panel. Photo credit: Andrew Zimolzak.

Physician Culture as a Barrier to Spread of Innovation

In the following panel moderated by critical care physician Leo Anthony Celi, MD, MS, MPH, fellow intensivists Djillali Annane, MD, PhD, Peter Clardy, MD and Taylor Thompson, MD reflected on the barriers presented by the current clinician culture toward the goal of data-driven innovation in medicine (Figure 4). The panelists observed that historically, EBM was perceived to be incompatible with well-established observational trials and experience, perhaps instilling a residual degree of resistance. Consequently, echoing Safran's sentiments, it will be increasingly important that "big data" is understood as a complement to RCTs and (patho)physiologic studies. Furthermore, condensing and filtering the vast quantity of data to make it applicable at the bedside will be key to adoption. The specific inclusion of clinicians during the design process will help to deter the creation of tools that inundate staff with extraneous information and burdensome extra tasks. Likewise the incorporation of "big data" into medical education, in a way that students and resident trainees will be able to understand its importance in both everyday care and expediting research, is vital.

While the panel agreed that more evidence is required to determine whether big data can facilitate comparative effectiveness research, it was acknowledged that it is necessary to investigate this alternative since RCTs do not, and will not, provide answers to an important fraction of the decisions required on a daily basis. Scaling up RCTs to account for the

thousands of decisions each day is not feasible, so big data approaches may provide the most effective way to fill these gaps. For example, three groups currently leading clinical trials research in the analysis of fluid resuscitation in critically ill patients have collaborated to create a common database architecture to allow for individual patient meta-analysis and for these trials to be evaluated in aggregate by an external monitoring committee.

Figure 4. Physician culture panel. Photo credit: Andrew Zimolzak.

The Role of Industry in the Data Revolution in Health Care

There is concern that industry continues to view data as a potential source of revenue and would therefore be opposed to providing open access to what they consider to be proprietary data with business value. In the last panel discussion of the day, moderated by Ambar Bhattacharyya, MBA of Bessem Venture Partners, industry panelists Josh Gray, MBA of AthenaResearch, Enakshi Singh, MS of SAP, and Omar Badawi, PharmD, MPH of Philips Healthcare provided their insights on the topic.

They first addressed the issue of sharing databases freely, noting that concerns associated with data ownership are not restricted to industry. Similar problems and conflicts are observed among most stakeholders in health care data ownership: patients, hospitals, providers, payers, vendors, and academia. Generally speaking, industrial data owners want to protect their data from those who may use it competitively against them, share or sell the data to derive direct clinical value, or profit from possible insights. They also wish to avoid the overhead costs associated with sharing. They are interested in allowing society to leverage their data in order to make gains if, and only if, these other interests remain unaffected.

The costs for responsibly sharing secondary clinical data are not trivial. Although understanding the complexity of the data presents a significant challenge, understanding the workflow for entering data in the primary system is often even more complicated, requiring extensive support. Therefore, sharing secondary clinical data can be a costly initiative for industry, lowering its priority as a business objective. These challenges are further exacerbated when collaboration requires intellectual property agreements. Lack of an accepted standard practice for research agreements, coupled with an outdated patent system, creates barriers to collaboration that are rarely overcome. Many ideas for collaboration either take years to initiate or never come to fruition, due to challenges with developing de novo legal research agreements.

While industry and researchers are not philosophically opposed to sharing data to ensure reproducibility, protections from the aforementioned concerns are critical. Can the data be shared without the risk of lost intellectual property? If not, the incentives for innovation may be minimized. Who will bear the

costs for ensuring that the replicating team fully understands the nuances of the data? Who will prevent competitors or others with malicious intent from inappropriately labeling valid research as "junk science"? Such underhanded interventions could introduce confusion around valid earlier findings and unfairly distract and denigrate the primary researchers.

Ultimately, there is a growing sense that data will become less of a commodity over time if governments continue to support the development and maintenance of open access research networks. As the scale and quality of these surpass those of privately owned databases, society will benefit as the obstacles to collaboration and the value of retaining private ownership diminish.

The Unreasonable Effectiveness of Data

Peter Szolovits, PhD from the MIT Computer Science and Artificial Intelligence Lab, highlighted how big data can often trump good, but smaller, data. Researchers at Google have been making a similar argument from a decade-long experience with natural language processing, showing that for some important tasks an order of magnitude growth in the size of a dataset leads to improvements in performance that can overshadow improvements in modeling technique [15]. They have also argued that discarding rare events is a bad idea because although these may be individually rare, they could prove to be significant later when examined on a much larger scale.

In the clinical world, patient state depends on complex pathophysiology dictated by genetic predispositions, environmental exposures, treatments, and numerous other factors. While, there are many potential ways to formulate clinical outcomes into complex statistical models, it is often the simple models that give the best, and most interpretable, results. Some clinicians and epidemiologists have already used large sources of observational data to improve clinical practice, especially in identifying drug side-effects, for example, for rosiglitazone [16], and rofecoxib [17]. Cox proportional hazard, naïve Bayes, linear and logistic regression, and similar models can use aggregated variables to summarize dynamic variation without adding additional complexity.

The Story of MIMIC: Open-Access Critical Care Data

Since researchers who seek to create new clinical knowledge and tools are dependent upon the availability of relevant data, restricting access to data introduces barriers that stifle research progress. This simple principle has been at the heart of the research of Roger Mark, MD, PhD since the 1980s, a time when his work was focused on developing real-time arrhythmia analysis tools for use in patient monitoring.

Like today, the norm for researchers in the 1980s was to privately maintain closed databases for their own benefit. So when Mark's team needed data, they began the painstaking work of creating their own resource, collecting electrocardiograms from patients at Boston's BIDMC and in the process, adding over 100,000 annotations. Breaking from tradition, they openly shared the dataset, reasoning that the more people who analyzed it, the better the overall understanding of arrhythmias would become. This dataset become known as the MIT-Beth Israel Hospital (MIT-BIH) Arrhythmia Database [18].

The consequences were far-reaching. Not only did the MIT-BIH Database stimulate research interest, it generated beneficial competition and became a shared resource for evaluating algorithms. Researchers competed to see whose work performed best on the standard data, eventually leading to the database becoming part of a federal requirement for evaluation of commercial algorithms. This success led the team to develop further resources unique in their openness, including PhysioNet, a platform for open physiologic data, and MIMIC II, a rich database of critical care data.

PhysioNet has over 50,000 registered users in over 120 countries and international recognition for accelerating the pace of discovery [19]. Mark attributes much of PhysioNet's success to the progressive mindset of the participating collaborators. Success has required not only funding, but also a collaborative approach among partnering clinicians, researchers, hospital technologists, and local ethics committees. Participation of commercial partners was also required, and obtained, in order to decrypt the proprietary data formats output by their monitoring systems.

Reproducibility of research and open data are increasingly getting the attention they deserve, but changing practice requires support at all levels [8]. For open technology to be embraced, funders must recognize the added value from a robust database infrastructure and allocate funds accordingly. Researchers too must embrace open approaches that perhaps challenge some of the underlying career reward systems. With changing attitudes, and by engaging the creative energy of the worldwide research community, Mark's hope is that MIMIC will become a multinational resource leading to the generation of new knowledge and new tools.

Opportunities and Challenges in Wearable Sensor Datasets

The ability to create and capture data is exploding and offers huge potential for health organizations around the world to save both lives and scarce resources. Yadid Ayzenberg, PhD discussed the "Opportunities and Challenges in Wearable Sensor Data" in his talk, focusing on how the combination of wearable technology and the near ubiquitous access to mobile phones have the potential to address some of the challenges in health care. Examples include the works of Poh et al [20] and Sano and Picard [21], which used a wrist-worn electrodermal activity and accelerometry biosensor for detection of convulsive seizures and sleep stages.

Wearable technologies provide a way to transition from a traditional aperiodic "snapshot" monitoring approach to a continuous and longitudinal monitoring paradigm, increase patients' engagement in their care, and facilitate doctor-patient interactions. Already massive amounts of personal health data are being generated through consumer devices such as mobile

phones and wristbands that monitor sleeping patterns, exercise, stress, calorie consumption and more. In most instances, however, the data are stored on a per-device basis, and there are unsolved issues concerning data management, ownership, privacy, and misuse. The noise and artifacts in the data measured by wearable sensors also present an important challenge. New analytic methods that transform "dirty data" into good quality data are needed.

Big Data, Genomics, and Public Health

According to Winston Hide, PhD, the promise of a new economy based on data-driven discovery and decision-making is also motivating his own field of genomics. Advances in genome sequencing technology will allow the cost of sequencing a genome to be less than $100 in the near future. Consequently, it is estimated that by 2015, one million genomes will be digitally available with high expectations for public health benefits. However, Hide cautions that there is still a "genome variants" problem to be solved. This was exemplified by the case of Kira Peikoff who was predicted to have a 20% above average risk of developing psoriasis by one commercial sequencing product, while predicted to have a 2% below average risk for the same condition by another [22].

Variations in the reporting of genomic characteristics have two potential root causes. First, there is a sampling problem caused by the use of different sequencing technologies, which introduces errors in evaluating genome assembly. Second, there is an interpretation problem in defining the role of genes in disease which compromises the prediction of clinical outcomes as determined by the single-nucleotide polymorphism analysis tools.

The availability of millions of genomes will allow completion of the catalogue of genes associated with a particular disease in genome-wide association studies. However, Hide maintains that finding clear drug targets requires the creation of an evolving catalogue of functions which would interpret complex gene pathways [23], and the selection of cohorts that would not only depend on ethnicity (the classic phenotype), but also on physiological and even molecular differences.

The application of genomic tests to public health will contribute to the transformation of physicians into data-centric specialists and pave the way for "precision medicine" [24]. This challenging way of delivering health care calls for new strategies and tactics for translating research into clinical practice. These will likely include the creation of open-access genomic and clinical databases, use of a common scientific language [25] and (open) data access tools. Regulatory bodies, such as the Food and Drug Administration, will have a role in guaranteeing the standardization and reliability of diagnostics based on genetic tests. These tools must guarantee the reproducibility [8] of discovered genome signals and contribute to the improvement of online platforms that map genetic features to diseases and their treatment (eg, Cancer Genome Atlas; PharmGKB).

The Pitfalls and Potential of Big Data in Health Care

Proponents of big data have made grandiose claims of expanding human knowledge by orders of magnitude through empirical analysis and data mining, but as Stanford professor John Ioannidis, MD, PhD says, "with big data comes big problems". Ioannidis discussed the darker side of data analysis, in which bias has led a large proportion of published medical science to come to the wrong conclusion. Author of the most downloaded paper in PLOS Medicine, "Why Most Published Research Findings Are False" [26], Ioannidis argued that most statistically significant results are likely to be false positives. For example, using the national drug and cancer registry database of Sweden, Ioannidis and colleagues found that almost one third of the 560 medications evaluated in isolation were associated with a higher cancer risk.

As Ioannidis highlights, the issue is not in the quantity of data we have. Increasing sample sizes is a huge boon to the medical field. The issue resides in a lack of transparency. When he reviews a paper published in a journal on a new dataset, his thoughts immediately drift to those studies that were not published. This is quantified in the so-called "vibration of effects", where depending on the confounding variables for which adjustments are made, completely opposite conclusions can be drawn. For vitamin E, for example, adjusting for a certain subset of confounders led to the conclusion that it increases the relative risk of mortality, whereas adjusting for another equally plausible set of confounders gave the opposite result, that is, a reduction in mortality risk. This may explain why 90% of effects in RCTs were lower in subsequent published trials [27].

Comparative Effectiveness Using Big Data

Limitations aside for now, ROS do provide an opportunity to conduct comparative effectiveness studies on research questions that would be unlikely to be examined by an RCT, or would be inherently biased if an RCT were conducted. The illustrative example presented by Una-May O'Reilly, PhD was the question of the potential benefit of diuretic use to accelerate removal of fluids given during resuscitation in ICU patients who have recovered from sepsis. Retrospective analysis would be easily marred by "selection bias". In fact, if patients are allocated (not randomized) to two groups, treatment and non-treatment, it is very likely that the allocation would be done on the basis of patient condition and biased by clinical severity resulting in unreliable results.

In a ROS, the data consists of a series of days during which the treatment was administered (D+) or not (D-). Because these decisions were being made on a daily basis, it is even harder to capture the covariance structure. O'Reilly refers to this as the "Non-Decision Day Dilemma". In order to deal with this, the covariance structure has to account for time-varying information with respect to a specific day. It is easy to take the treatment day as a reference and align all patients who received treatments with respect to this event (D+). For non-treated patients, aligning

time-series is more complicated as every day is essentially a "non-decision day". Considering every single day would result in a widely unbalanced dataset where the length-of-stay influences the individual contribution of each patient. To account for this, it is possible to randomly sample N negative days (D*-) and pair them up with the positive instances (D+) based on a statistical similarity criterion with respect to the time from admission. This is achieved by defining a propensity score for each patient for every day during their ICU stay. The propensity score thus enables appropriate cohort matching such that comparative effectiveness can be appropriately assessed.

This modest example illustrates the sort of robust and reliable statistical technique that evidence-based medicine requires. It can reduce sample noise and improve the reliability of conclusions, and it leads toward methodology standardization across studies. Beyond these local improvements, meta-studies will also be a requirement to validate any local finding. These are only possible with data sharing and open data initiatives such as the MIMIC-II initiative. One strength of this database is the dual culture and scientific activity it generates because data science can only fully benefit from collaboration between data scientists and domain experts (in this case, intensivist physicians).

Conclusions

Although the future of "big data" in health care remains unclear, its role will be undeniably important. This conference was effective in collating the broad range of perspectives on the many challenges facing EBM in the 21st Century. As several speakers suggested, one possible opportunity is to adopt a pragmatic approach to EBM, combining RCT and ROS. This combination may employ ROSs to fill the gaps where it is impractical, unlikely or impossible to conduct a RCT or to drive hypothesis generation for further RCT analysis.

It is also crucial to acknowledge that any ROS requires a multidisciplinary approach, integrating clinical knowledge with a broad range of data analytic skills ranging from biostatistics, machine learning, and signal processing to data mining. Encouraging a change in physician culture can likely be accomplished through updating education programs as well as by creating centers for excellence that can showcase the impact of ROS to the broader medical fraternity. These centers for excellence should host open, transparent, easily accessible data warehouses, which will facilitate study reproducibility and allow for a new wave of collaborative learning. Only by understanding the potential biases of any analysis, and fostering a system of normative data sharing, will the medical community be able to gain reliable knowledge from data, and produce research findings that do not turn out to be false.

Acknowledgments

The Organizing Committee would like to thank the speakers and panelists, the sponsors and all the attendees of the MIT Critical Data Conference and Marathon held in January 2014. The MIT Critical Data Marathon and Conference 2014 was sponsored by SAP, Philips Healthcare, Quanttus, Goodwin Procter and WilmerHale.

Conflicts of Interest

Omar Badawi is senior clinical scientist at Philips Healthcare. Christopher Moses is founder and CEO of Smart Scheduling, Inc. Louis Mayaud is research director of Mensia Technologies SA. Thomas Brennan is data scientist at AthenaHealth Research.

References

1. Celi LA, Mark RG, Stone DJ, Montgomery RA. "Big data" in the intensive care unit. Closing the data loop. Am J Respir Crit Care Med 2013 Jun 1;187(11):1157-1160. [doi: 10.1164/rccm.201212-2311ED] [Medline: 23725609]
2. Grande D, Mitra N, Shah A, Wan F, Asch DA. Public preferences about secondary uses of electronic health information. JAMA Intern Med 2013 Oct 28;173(19):1798-1806. [doi: 10.1001/jamainternmed.2013.9166] [Medline: 23958803]
3. McAfee A, Brynjolfsson E. Big data: the management revolution. Harv Bus Rev 2012 Oct;90(10):60-6, 68, 128. [Medline: 23074865]
4. Bourne PE. What Big Data means to me. J Am Med Inform Assoc 2014;21(2):194. [doi: 10.1136/amiajnl-2014-002651] [Medline: 24509599]
5. The MIMIC II project. Cambridge, MA: Massachusetts Institute of Technology URL: http://mimic.physionet.org [accessed 2014-08-11] [WebCite Cache ID 6RkH0iGr1]
6. Critical data: Empowering big data in critical care. URL: http://criticaldata.mit.edu/past-events/ [accessed 2014-08-20] [WebCite Cache ID 6Rxk41BBR]
7. Freedman DH. The Atlantic. 2010 Oct 4. Lies, damned lies, and medical science URL: http://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269/ [accessed 2014-08-11] [WebCite Cache ID 6RkHK88k4]
8. The Economist. London, UK; 2013 Oct 19. Trouble at the lab URL: http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble [accessed 2014-08-11] [WebCite Cache ID 6RkHeZm6r]
9. Moses C, Celi LA, Marshall J. Pharmacovigilance: an active surveillance system to proactively identify risks for adverse events. Popul Health Manag 2013 Jun;16(3):147-149. [doi: 10.1089/pop.2012.0100] [Medline: 23530466]

10. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Expert Panel. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007 Jan;14(1):1-9 [FREE Full text] [doi: 10.1197/jamia.M2273] [Medline: 17077452]
11. Geissbuhler A, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, et al. Trustworthy reuse of health data: a transnational perspective. Int J Med Inform 2013 Jan;82(1):1-9. [doi: 10.1016/j.ijmedinf.2012.11.003] [Medline: 23182430]
12. Cobb JP, Cairns CB, Bulger E, Wong HR, Parsons PE, Angus DC, et al. The United States critical illness and injury trials group: an introduction. J Trauma 2009 Aug;67(2 Suppl):S159-S160. [doi: 10.1097/TA.0b013e3181ad3473] [Medline: 19667851]
13. Reinhart CM, Rogoff KS. A Decade of Debt. Cambridge, MA: National Bureau of Economic Research; 2011. Working Paper 16827 URL: http://www.nber.org/papers/w16827.pdf [accessed 2014-08-11] [WebCite Cache ID 6RkHujrwl]
14. Herndon T, Ash M, Pollin R. Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff. Amherst, MA: Political Economy Research Institute; 2013. Working Paper Series No. 322 URL: http://www.peri.umass.edu/fileadmin/pdf/working_papers/working_papers_301-350/WP322.pdf [accessed 2014-08-11] [WebCite Cache ID 6RkI0b5Rf]
15. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. IEEE Intell Syst 2009;24(2). [doi: 10.1109/MIS.2009.36]
16. Brownstein JS, Murphy SN, Goldfine AB, Grant RW, Sordo M, Gainer V, et al. Rapid identification of myocardial infarction risk associated with diabetes medications using electronic medical records. Diabetes Care 2010 Mar;33(3):526-531 [FREE Full text] [doi: 10.2337/dc09-1506] [Medline: 20009093]
17. Brownstein JS, Sordo M, Kohane IS, Mandl KD. The tell-tale heart: population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial infarction. PLoS One 2007;2(9):e840 [FREE Full text] [doi: 10.1371/journal.pone.0000840] [Medline: 17786211]
18. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag 2001 May;20(3):45-50. [Medline: 11446209]
19. Kalil T, Green E. Whitehouse Office of Science and Technology Policy. Washington, DC Big Data is a Big Deal for Biomedical Research Internet URL: http://www.whitehouse.gov/blog/2013/04/23/big-data-big-deal-biomedical-research [accessed 2014-08-11] [WebCite Cache ID 6RkISsDjO]
20. Poh MZ, Loddenkemper T, Reinsberger C, Swenson NC, Goyal S, Sabtala MC, et al. Convulsive seizure detection using a wrist-worn electrodermal activity and accelerometry biosensor. Epilepsia 2012 May;53(5):e93-e97. [doi: 10.1111/j.1528-1167.2012.03444.x] [Medline: 22432935]
21. Sano A, Picard RW. Toward a taxonomy of autonomic sleep patterns with electrodermal activity. Conf Proc IEEE Eng Med Biol Soc 2011;2011:777-780. [doi: 10.1109/IEMBS.2011.6090178] [Medline: 22254426]
22. Peikoff K. The New York Times. I Had My DNA Picture Taken, With Varying Results URL: http://www.nytimes.com/2013/12/31/science/i-had-my-dna-picture-taken-with-varying-results.html?_r=0 [accessed 2014-08-11] [WebCite Cache ID 6RkIrPHId]
23. Lander AD. The edges of understanding. BMC Biol 2010;8:40 [FREE Full text] [doi: 10.1186/1741-7007-8-40] [Medline: 20385033]
24. Peterson TA, Doughty E, Kann MG. Towards precision medicine: advances in computational approaches for the analysis of human variants. J Mol Biol 2013 Nov 1;425(21):4047-4063. [doi: 10.1016/j.jmb.2013.08.008] [Medline: 23962656]
25. Sansone SA, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, et al. Toward interoperable bioscience data. Nat Genet 2012 Feb;44(2):121-126 [FREE Full text] [doi: 10.1038/ng.1054] [Medline: 22281772]
26. Ioannidis JP. Why most published research findings are false. PLoS Med 2005 Aug;2(8):e124 [FREE Full text] [doi: 10.1371/journal.pmed.0020124] [Medline: 16060722]
27. Pereira TV, Horwitz RI, Ioannidis JP. Empirical evaluation of very large treatment effects of medical interventions. JAMA 2012 Oct 24;308(16):1676-1684. [doi: 10.1001/jama.2012.13444] [Medline: 23093165]

Abbreviations

ACO: accountable care organization
BIDMC: Beth Israel Deaconess Medical Center
CMS: Center for Medicare and Medicaid
EBM: evidence-based medicine
EMR: electronic medical record
ICU: intensive care unit
MIMIC: Multiparameter Intelligent Monitoring in Intensive Care
MIT: Massachusetts Institute of Technology
MIT-BIH: Massachusetts Institute of Technology – Beth Israel Hospital
RCT: randomized controlled trial
ROS: retrospective observational study

Edited by G Eysenbach; submitted 03.04.14; peer-reviewed by D Maslove, L Toldo, S Seevanayanagam; comments to author 15.07.14; revised version received 24.07.14; accepted 25.07.14; published 22.08.14.

Please cite as:
Badawi O, Brennan T, Celi LA, Feng M, Ghassemi M, Ippolito A, Johnson A, Mark RG, Mayaud L, Moody G, Moses C, Naumann T, Nikore V, Pimentel M, Pollard TJ, Santos M, Stone DJ, Zimolzak A, MIT Critical Data Conference 2014 Organizing Committee
Making Big Data Useful for Health Care: A Summary of the Inaugural MIT Critical Data Conference
JMIR Med Inform 2014;2(2):e22
URL: http://medinform.jmir.org/2014/2/e22/
doi:10.2196/medinform.3447
PMID:

Original Paper

A Validation of an Intelligent Decision-Making Support System for the Nutrition Diagnosis of Bariatric Surgery Patients

Magda RR Cruz[1], MSc; Cristina Martins[2]; João Dias[3], DSc (Tech); José S Pinto[4], DSc (Tech)

[1]Pontifical Catholic University of Paraná (PUCPR), Curitiba, Brazil

[2]Cristina Martins Institute and Kidney Foundation, Curitiba, Brazil

[3]Department of Electrical Engineering, Federal University of Parana, Curitiba, Brazil

[4]Science, Management, and Information Technology, Federal University of Paraná, Curitiba, Brazil

Corresponding Author:
Magda RR Cruz, MSc
Pontifical Catholic University of Paraná (PUCPR)
St Imaculada Conceição, 1155
Curitiba, 80215901
Brazil
Phone: 55 41 32426543
Fax: 55 41 32426543
Email: nutriclinmagda@bol.com.br

Abstract

Background: Bariatric surgery is an important method for treatment of morbid obesity. It is known that significant nutritional deficiencies might occur after surgery, such as, calorie-protein malnutrition, iron deficiency anemia, and lack of vitamin B12, thiamine, and folic acid.

Objective: The objective of our study was to validate a computerized intelligent decision support system that suggests nutritional diagnoses of patients submitted to bariatric surgery.

Methods: There were fifteen clinical cases that were developed and sent to three dietitians in order to evaluate and define a nutritional diagnosis. After this step, the cases were sent to four bariatric surgery expert dietitians who were aiming to collaborate on a gold standard. The nutritional diagnosis was to be defined individually, and any disagreements were solved through a consensus. The final result was used as the gold standard. Bayesian networks were used to implement the system, and database training was done with Shell Netica. For the system validation, a similar answer rate was calculated, as well as the specificity and sensibility. Receiver operating characteristic (ROC) curves were projected to each nutritional diagnosis.

Results: Among the four experts, the rate of similar answers found was 80% (48/60) to 93% (56/60), depending on the nutritional diagnosis. The rate of similar answers of the system, compared to the gold standard, was 100% (60/60). The system sensibility and specificity were 95.0%. The ROC curves projection showed that the system was able to represent the expert knowledge (gold standard), and to help them in their daily tasks.

Conclusions: The system that was developed was validated to be used by health care professionals for decision-making support in their nutritional diagnosis of patients submitted to bariatric surgery.

(JMIR Med Inform 2014;2(2):e8) doi:10.2196/medinform.2984

KEYWORDS

bariatric surgery; nutrition diagnosis; artificial intelligence; Bayesian networks; decision-making; support system

Introduction

Nutrition and Bariatric Surgery

Morbid obesity causes a number of health issues, explaining why, in certain situations, some aggressive treatments may be used, for instance, bariatric surgery. The surgical procedure is indicated when the patient presents a body mass index over 40 kg/m^2, or when it is situated between 35 and 40 kg/m^2 and also presents some associated disease, such as, diabetes, dyslipidemias, cardiovascular and cerebrovascular diseases, sleep apnea, joint disease, and orthopedic disease, among others [1]. It is estimated that one million bariatric surgeries will be performed in the next few years in the United States alone [2].

Therefore, the concern related to the nutritional changes in the long term in these patients is highly important [3-7]. Furthermore, the need for individualized management of patients with obesity is evident [3-7]. Thus, the health professional concern related to some special nutritional care is comprehensible, particularly in relation to eating in the pre and post operatory in bariatric surgeries.

Some of the most common nutritional deficiencies include iron, vitamin B12, folate, thiamine, and protein after bariatric surgery [2,5,8-11]. Severe consequences can be expected when they are not prevented or treated early.

The Nutrition Care Process

The Nutrition Care Process consists of four steps: (1) nutrition assessment, (2) nutrition diagnosis, (3) nutrition intervention, and (4) nutrition monitoring and evaluation. The nutrition diagnosis, the second step of the Nutrition Care Process, is the identification and record that describes an actual occurrence, risk of, or potential for developing a nutritional problem [12].

The results from the use of this technology, which are achieved by computing beyond the nutritional science knowledge, are important in order to help in detecting nutritional deficiencies. The information technology in the field of health has tools and instruments that may support the administrative organization in patient service. These tools and instruments are able to capture, store, and process information, and may offer some diagnosis suggestions, therapeutic orientation, and access to information [13]. The specialized systems are very helpful for the health professionals. In particular, there is the so-called Decision Support System (DSS).

These programs are used to help the professionals to define the diagnosis through artificial intelligence. A Bayesian network (BN) is the technique used in the formulation of DSS. It is able to represent the uncertainty in knowledge through the Bayes' Theorem. In this case, the necessary data for the model is collected through published statistical studies and/or through specialist consultation [14]. The Bayes' Theorem calculates the probabilities in each diagnosis, given a set of pre existing information [14]. The fact that it can work with uncertainty through probability makes it the most significant technique to be used in the health field.

Aim of the Study

The aim in this study is to validate a DSS that will help the nutritional diagnosis for bariatric surgery patients through the development of a protocol created by experts in the field, given the large number of surgeries, the long term nutritional risks, the small amount of specialists in the field, and the absence of a specific computer system.

Methods

The Selection Process

The prevalence of each nutritional diagnosis has different probabilities, depending on the bariatric surgery technique used. Therefore, only patients submitted to the surgical technique Roux-en-Y gastric bypass were selected for this study. These diagnoses are currently considered the gold standards [15].

The First Stage

The first stage of the study comprised the knowledge base building. There were two resources that were used in order to do so: (1) scientific studies published in internationally recognized journals, in addition to important studies in the fields of nutrition and medicine; and (2) consultations with nutrition specialists. From these sources, the major nutritional deficiencies presented in the post operatory were verified [1,2,9-11], the average weight loss found in patients was noted [16], the main signs and symptoms in patients were described [8,11,17], and the definitions of the techniques used in the nutritional assessment were identified [18,19].

The results from this stage indicated that a specialized module of nutritional diagnosis should consider gender, age, surgery time, biochemical markers (hemoglobin, hematocrit, mean corpuscular volume, serum albumin, ferritin, vitamin B12 and folic acid), food intake, and physical signs and symptoms of nutrient deficiency. This study opted for classifying them as high, low, or normal, according to the usual standard references, due to a wide range of techniques to measure the selected biochemical markers. The analysis of the number of food portions consumed for the food intake evaluation was based on the Food Guide Pyramid [19]. The reference was 1600 kcal, which is the minimum amount recommended for a suitable macro and micronutrients intake. The physical signs (hair loss, changes in nails and skin, paleness) and symptoms (weakness, paresthesia, vomiting, diarrhea, blood loss) are derived from the subjectivity of professionals who qualify the information before it is used by the system. Because the data on dietary intake and the signs and symptoms are subjective, BNs have been selected for the representation of knowledge. The technique considers the evidence presented for the calculation of the disease probability in case it happens, and allows that the subjectivity or the uncertainty element of information be considered. Last, the standard nutritional diagnoses were protein-energy malnutrition, iron deficiency anemia, vitamin B12 deficiency, folic acid deficiency, and thiamine deficiency. Additionally, tools to identify risks to develop all these deficiencies were created.

From the tools mentioned above, a study of the variables was carried on, considering each nutritional diagnosis for each patient. For instance, for a patient with iron deficiency, all the signs, symptoms, dietary intake, and biochemical markers indicated from the literature were analyzed. All of the information that either caused doubts or did not help in the diagnosis conclusion were excluded for not being considered decisive in the decision support. In other words, only the variables that influenced in the diagnosis decision were kept in the study.

After assembling the qualitative part of the network (inclusive and exclusive definition of variables), probabilistic values were assigned for each of them, as described in the literature. Thus, the quantitative part of the network was originated. As there was no availability of a database containing all the variables and attributes required to work, the use of literature and discussion with experts were chosen. For each nutritional diagnosis, the probability of the event in the presence or absence

of the disease, or the risk of the development of each one of the variables, was considered.

The Technology Used

The technique implementation of the BN was performed with the aid of Shell Netica. It has the infrastructure to develop expert systems within a pre built interface. The program Netica is composed by Netica Application and the Netica Application program interface (API). The Netica Application is a graphical interface that permits you to view the knowledge base in a network. The Netica API is the library of the program, written in C language, which is available on a website [20] on the Internet.

Preliminary System Evaluation

In the first step of the nutritional diagnosis support system validation, fifteen case studies were developed and elaborated on by two nutritionists; one was an expert in morbid obesity, and the other one was not. All the case studies were sent to four expert nutritionists in the field of nutrition in order to get evaluations and diagnosis reports from them. A standard diagnosis list was attached to the case studies. It was also requested that the evaluators suggest changes in the developed clinical cases and in the diagnostic proposal. The four experts' answers were compared to those given by the system, and the experts' answers were revised based on this evaluation. Thereby, a proposal for the nutritional diagnosis support system was presented called DSS Diagnosis Nutrition 1. This contained the case studies reviewed, according to the nutritionists' opinion.

Gold Standard Development

The experts were selected for the gold standard development based on: (1) nutrition studies background, (2) over two years as a member of the multidisciplinary team for the treatment of patients submitted to the obesity surgery, and (3) if they have followed more than 300 patients in the post operatory. There were four experts that were selected according to these criteria. They received the fifteen case studies revised, and had to send diagnosis reports for each of them. The experts' reports were compared among themselves. The disagreements were solved through consensus among the experts, resulting in the gold standard. This standard aimed to evaluate the system performance.

System Validation Technique

The following analyses were performed for the final system validation: (1) comparison between the four experts' success rates, the gold standard success rate, and between the system and the gold standard; (2) calculation of sensibility and specificity for each nutritional diagnosis; and (3) the receiver operator characteristic (ROC) curve construction for each diagnosis.

All the ethical principles in the Helsinki Declaration (2000) [21] were respected during the development of this study. There was no direct participation of human beings.

Results

Success Rate Between the Experts and the Gold Standard

The qualitative part of the BN was done considering the interrelation among the nutritional diagnosis and the signs, symptoms, food intake, and biochemical markers. As a result, five subnets were obtained, each one of them featuring a nutritional diagnosis: (1) vitamin B12 deficiency, (2) thiamine deficiency, (3) folic acid deficiency, (4) iron deficiency, and (5) malnutrition. The health professional could classify the patients' diagnosis as "present", "absent", or "in risk" of developing it. At the same time, the four experts selected to build the gold standard diagnosis were questioned after assessing the fifteen clinical cases sent to them. That originated 60 answers per nutritional diagnosis. The experts' diagnoses were compared to the gold standard, creating the experts assertiviness rate related to the gold standard (Table 1).

Table 1. Success rate for nutritional diagnosis between the four experts and the gold standard/ BN algorithm.

Cases	Iron deficiency anemia	Folate deficiency	B12 deficiency	Thiamine deficiency	Malnutrition
Number of success/total	54/60	56/60	48/60	52/60	55/60
Assertiveness (%)	90	93	80	87	92
Standard deviation	23	11	25	23	15

Expert Disagreements

Vitamin B12 and thiamine deficiencies were the diagnoses that most presented disagreements among experts, followed by iron deficiency anemia. The values in Table 1 were 48, 52, and 54 assertiveness respectively, in a sample of 60 cases. In other words, the result was higher than that found among the four experts, which presented a variation between 80% (48/60) and 93% (56/60), according to the diagnosis. That showed that even though there are criteria for each nutritional diagnosis, the individual interpretation could make the task difficult.

The answers reported by the four experts were analyzed individually, causing greater disagreement in the definition of the diagnosis of the problem or presence of risk, thus, reinforcing the usefulness of the system to aid the diagnosis, either confirming the professional hypothesis or warning them of the disease risk.

System Assessment in Relation to the Gold Standard

Success Rate of the System

The diagnoses reports from the system were compared to the reports from the gold standard in order to assess the performance of the system. The success rate found was 100% (60/60) for the

case reports diagnosed. Taking as an example the first clinical case presented to the experts, the following situation was observed, the gold standard detected risk to the development of iron deficiency anemia, folic acid deficiency, thiamine deficiency, and malnutrition. None of the diagnoses were confirmed, and the presence or the risk of development of vitamin B12 deficiency was rejected. When the same data was input to the system, this presented values higher than 70.0% (70/100) of risk of development of folic acid deficiency, of vitamin B1 deficiency, and of malnutrition. The values were 100.0% (100/100) for iron deficiency anemia. The vitamin B12 deficiency, which was a diagnosis rejected by the gold standard, presented 0.11% (.11/100) chance of confirmation and 33.5% (33.5/100) of risk of development. The same happened to the other cases, thus proving that there was agreement between the reports provided by the system and the gold standard.

The percentage for the diagnosis and for the risk of development was very close in some cases, when each case was analyzed individually. For instance, in case 3 the patient presented 42.3% (42.3/100) of probability of confirmation of the diagnosis for anemia, and 56.3% (56.3/100) of probability of risk of development of anemia. The gold standard classified the patient as in risk of anemia, agreeing with the system.

Sensibility and Specificity for Each Nutritional Diagnosis and the Construction of the Receiver Operator Characteristic Curve

The Medicalc was used for the analysis of sensibility and specificity, determining the ability to discriminate among diagnoses through the ROC curve. There was a comparison between the reports from the system and those from the gold standard. The results are in Table 2.

Table 2. BN diagnosis test results.

Test assessment	Presence of risk or diagnosis x absence				
	Iron	Folic acid	B12	Thiamine	Malnutrition
Sensibility %	95.0	95.0	95.0	95.0	95.0
Specificity %	95.0	95.0	95.0	95.0	95.0
Area below the ROC curve	0.893	1.0	1.0	0.982	1.0
Standard error	0.088	0.0	0.0	0.038	0.0
95% confidence interval	0.627 - 0.986	0.78 - 1.0	0.78 - 1.0	0.751 - 1.0	0.78 - 1.0

Results of the Comparison

It was observed that the specificity and the sensibility of the system presented high levels (95.0%) for all the diagnoses. The results reflect the validation of its use. In other words, the system is able to represent the gold standard. Besides that, the confidence interval was well established and the standard error was low (0.0-0.088). The results also confirmed the agreement between the gold standard and the system.

Regarding the ROC curve, it was observed that for the folic acid deficiency, vitamin B12, and malnutrition, the system presented maximum performance (1.0).

The results found for iron deficiency anemia and for thiamine deficiency also indicated a good performance of the system. However, there was a small deviation in its projection, which represents the possibility of disagreement between the system and the gold standard. In the end, the analysis of the data showed in the ROC curve concluded that the system presented a good performance in the definition of each diagnosis, thus being able to be used in the aid of health professionals.

Discussion

Expert Diagnoses

The success rate from the developed system was higher than that found among the four experts. This result reflects the difficulty in the diagnosis definition by the specialists. The fact is comprehensible since the definition of a diagnosis involves different information, previous experiences, and many times, the use of common sense and intuition. The mental mechanisms and the processes of thinking used by a specialist to arrive at a diagnosis are still poorly understood. Many times there is a lack of consensus among specialists, in some fields [22,23]. Furthermore, as the nutritional following-up of patients who were submitted to a bariatric surgery is still something recent, the disagreement among professionals may be common. In this study, the development of a gold standard by nutrition specialists was essential. Not only due to the need of a reference, but also for creating a discussion and reflection regarding the diagnosis that each specialist had previously established. This discussion confirmed the need of a system that makes the professional think about other possibilities, before the final nutritional diagnosis.

Decision Support Systems

There are not any other intelligent DSSs that have been developed specifically for the nutritional monitoring of patients undergoing bariatric surgery known by this group. Because of that, there was no chance of our system being compared to any other similar systems. Quick Medical Reference System is a system that helps in the diagnosis of many fields in medicine. It presents a success rate of 85% [24]. Our system presents a success rate of 100% (60/60) for the case reports diagnosed. Therefore, its good performance is confirmed as well as its indication of use.

The developed system in this study presents the possibility of working with the probability of risk/disease, versus the absence of nutritional risk, and enables the health professional not only to detect diseases, but also to detect the risk of developing them.

Thus, it increases the possibility of prevention, of early treatment, or even a specific follow-up, therefore preventing a disease from progressing to more serious stages. Thus, it is expected that the system assists in the patients' follow-ups, not only suggesting the nutritional diagnosis, but also preventing the major deficiencies that can occur post operatively in bariatric surgery.

Another extremely important factor that should be considered is the possibility of changing or adding variables to the system in the future, as the developments of new studies and the experts'/users' opinions occur. This aspect facilitates the maintenance and improvement of the system's performance. The inclusion of data to assist in the diagnosis of other nutritional deficiencies, and that are currently being researched, may enrich the system in the future. This is the situation of osteoporosis, which may occur in the late post operative period, or even the zinc deficiency, which is often mentioned, but rarely diagnosed in clinical practice.

Conclusions

This study enabled the validation of a DSS to assist the health professional in the nutritional monitoring of patients submitted to bariatric surgery.

The aim has been achieved since the system was able to duplicate the reports issued by the gold standard, both in the presence of disease and in the risk of developing it. The construction of an elaborate knowledge base proves to be essential in obtaining results. The result of showing the probabilities of the patient having the disease or the risk of developing it, rather than just issuing reports with categorical outcomes ("yes" or "no"), increases the freedom of the professional making the decision. In other words, it does not make the diagnosis authoritative, but suggestive.

It can be affirmed, through this study, that BN is an effective tool in the duplication of expert knowledge when there are several factors with different probabilities of occurrence involved in the definition of a diagnosis. Therefore, its use is indicated in health care.

In conclusion, the information gathered while developing DSSs for nutritional diagnosis can facilitate health professionals' tasks. The goal is not to replace professional work, but to help decision making. The intention is not to solve clinical cases, but to instigate critical thinking before the diagnostic decision.

Acknowledgments

This was the master's thesis from Magda Rosa Ramos da Cruz, for Catholic University of Paraná (PUCPR), directed by Professor João da Silva Dias, and codirected by Professor Cristina Martins, both from PUCPR.

Conflicts of Interest

None declared.

References

1. Kolanowski J. Surgical treatment for morbid obesity. Br Med Bull 1997;53(2):433-444 [FREE Full text] [Medline: 9246844]
2. Fujioka K. Diabetes Care. 2005. Follow-up of nutritional and metabolic problems after bariatric surgery URL: http://care.diabetesjournals.org/content/28/2/481.full.pdf [accessed 2014-04-18] [WebCite Cache ID 6OxZFy0AE]
3. Decker GA, Swain JM, Crowell MD, Scolapio JS. Gastrointestinal and nutritional complications after bariatric surgery. Am J Gastroenterol 2007 Nov;102(11):2571-2580. [doi: 10.1111/j.1572-0241.2007.01421.x] [Medline: 17640325]
4. Lemieux S, Mongeau L, Paquette M, Lberge S, Lachance B. Health Canada's new guidelines for body weight classification in adults: Challenges and concerns. Canadian Medical Association Journal 2004;171:1361-1363. [doi: 10.1503/cmaj.1032012]
5. Harbottle L. Audit of nutritional and dietary outcomes of bariatric surgery patients. Obes Rev 2011 Mar;12(3):198-204. [doi: 10.1111/j.1467-789X.2010.00737.x] [Medline: 20406412]
6. Cruz MRR, Morimoto IMI. Intervenção nutricional no tratamento cirúrgico da obesidade mórbida: Resultados de um protocolo diferenciado. Rev. Nutr 2004 Jun;17(2):263-272. [doi: 10.1590/S1415-52732004000200013]
7. Shikora SA. Techniques and procedures: Surgical treatment for severe obesity: The state-of-the-art for the new millennium. Nutrition in Clinical Practice 2000 Feb 01;15(1):13-22. [doi: 10.1177/088453360001500104]
8. Chaves LCL, Faintuch J, Kahwage S, Alencar FA. Revista Brasileira de Nutrição Clínica. 2002. Complicação pouco relatada em obesos mórbidos: Polineuropatia relacionada à hipovitaminose B1 / A rarely reported complication in morbid obesity - vitamin B1 - related polyneuropathy URL: http://bases.bireme.br/cgi-bin/wxislind.exe/iah/online/?IsisScript=iah/iah.xis&src=google&base=LILACS&lang=p&nextAction=lnk&exprSearch=316052&indexSearch=ID [WebCite Cache ID 6OxaOixDK]
9. Faintuch J, Matsuda M, Cruz ME, Silva MM, Teivelis MP, Garrido AB, et al. Severe protein-calorie malnutrition after bariatric procedures. Obes Surg 2004 Feb;14(2):175-181. [doi: 10.1381/096089204322857528] [Medline: 15018745]
10. Halverson JD. Micronutrient deficiencies after gastric bypass for morbid obesity. Am Surg 1986 Nov;52(11):594-598. [Medline: 3777703]
11. Alvarez-Leite JI. Nutrient deficiencies secondary to bariatric surgery. Curr Opin Clin Nutr Metab Care 2004 Sep;7(5):569-575. [Medline: 15295278]

12. Lacey K, Pritchett E. Nutrition Care Process and Model: ADA adopts road map to quality care and outcomes management. J Am Diet Assoc 2003 Aug;103(8):1061-1072. [doi: 10.1053/jada.2003.50564] [Medline: 12891159]
13. Glaser J, Douglas EH, Gregory D, Kristin MB. Advancing personalized health care through health information technology: An update from the American health information community's personalized health care. J. Am. Med. Inform. Assoc 2008 Jul;15(4):391-396. [doi: 10.1197/jamia.M2718]
14. Linda CVG. The Computer Journal. 1996. Bayesian belief networks: Odds and ends URL: http://www.cs.uu.nl/research/techreps/repo/CS-1996/1996-14.pdf [accessed 2014-04-19] [WebCite Cache ID 6OyvnXT1T]
15. Faintuch J, Oliveira CPMS, Rscovski A, Matsuda M, Bresciani JC. Revista Brasileira de Nutrição Clínica. 2003. Consideracoes nutricionais sobre a cirurgia bariatrica URL: http://www.sbnpe.com.br/consideracoes-nutricionais-sobre-a-cirurgia-bariatrica [accessed 2014-04-20] [WebCite Cache ID 6Oyw0Egze]
16. Faria OP, Pereira V, Gangoni CMC, Lins R, Lette S, Rassi S. Boletim de Cirurgia da Obesidade. 2001. Obesos mórbidos tratados com gastroplastia redutora com bypass gástrico em y de roux: Análise de 160 pacientes URL: http://bases.bireme.br/cgi-bin/wxislind.exe/iah/online/?IsisScript=iah/iah.xis&src=google&base=LILACS&lang=p&nextAction=lnk&exprSearch=356402&indexSearch=ID [WebCite Cache ID 6OyyC0jlm]
17. Berger JR. The neurological complications of bariatric surgery. Arch Neurol 2004 Aug;61(8):1185-1189. [doi: 10.1001/archneur.61.8.1185] [Medline: 15313834]
18. National Institutes of Health. NIH Publication. 2000. The practical guide: Identification, evaluation, and treatment of overweight and obesity in adults URL: http://www.nhlbi.nih.gov/guidelines/obesity/prctgd_c.pdf [accessed 2014-04-19] [WebCite Cache ID 6Oyz1w0Nd]
19. United States Departament of Agriculture - Center for Nutrition Policy and Promotion. 2014. My plate URL: http://www.cnpp.usda.gov/MyPlate.htm [accessed 2014-04-17] [WebCite Cache ID 6Ovv07QxB]
20. Anonymous. Norsys software corp. URL: http://www.norsys.com/ [accessed 2014-04-18] [WebCite Cache ID 6OxVMGh1s]
21. Declaration of Helsinki. 2000. WMA Declaration of Helsinki - Ethical principles for medical research involving human subjects URL: http://www.wma.net/en/30publications/10policies/b3/ [accessed 2014-04-18] [WebCite Cache ID 6Ovv6mLeN]
22. Javitt J. Computer-aided diagnosis and decision making. In: Computers in medicine: applications and possibilities. Philadelphia: Saunders; 1986.
23. Szolovits P. Artificial intelligence in medicine. Boulder, Colo: Published by Westview Press for the American Association for the Advancement of Science; 1982.
24. Widman LE. Informática Médica. 1998. Sistemas especialistas em medicina URL: http://www.informaticamedica.org.br/informaticamedica/n0105/widman.htm [accessed 2014-04-20] [WebCite Cache ID 6Oz0ETgmd]

Abbreviations

API: application program interface
BN: Bayesian networks
DSS: Decision Support System
PUCPR: Catholic University of Paraná
ROC: receiver operator characteristic

Edited by G Eysenbach; submitted 26.09.13; peer-reviewed by E Borsato, B Davy, D Steinberg; comments to author 22.10.13; revised version received 16.12.13; accepted 30.03.14; published 08.07.14.

Please cite as:
Cruz MRR, Martins C, Dias J, Pinto JS
A Validation of an Intelligent Decision-Making Support System for the Nutrition Diagnosis of Bariatric Surgery Patients
JMIR Med Inform 2014;2(2):e8
URL: http://medinform.jmir.org/2014/2/e8/
doi:10.2196/medinform.2984
PMID:25601419

Original Paper

Exploring a Clinically Friendly Web-Based Approach to Clinical Decision Support Linked to the Electronic Health Record: Design Philosophy, Prototype Implementation, and Framework for Assessment

Perry Miller[1,2,3], MD, PhD; Michael Phipps[4,5], MD; Sharmila Chatterjee[6], MD, MPH; Nallakkandi Rajeevan[2], PhD; Forrest Levin[1], MS; Sandra Frawley[2], PhD; Hajime Tokuno[1,7], MD

[1]VA Connecticut Healthcare System, West Haven, CT, United States
[2]Center for Medical Informatics, Yale University School of Medicine, New Haven, CT, United States
[3]Department of Anesthesiology, Yale University School of Medicine, New Haven, CT, United States
[4]Baltimore VA Medical Center, Baltimore, MD, United States
[5]Department of Neurology, University of Maryland School of Medicine, Baltimore, MD, United States
[6]Department of Medicine, Yale University School of Medicine, New Haven, CT, United States
[7]Department of Neurology, Yale University School of Medicine, New Haven, CT, United States

Corresponding Author:
Perry Miller, MD, PhD
Center for Medical Informatics
Yale University School of Medicine
300 George Street, Suite 501
New Haven, CT, 06511
United States
Phone: 1 203 737 2903
Fax: 1 203 737 5708
Email: perry.miller@yale.edu

Abstract

Background: Computer-based clinical decision support (CDS) is an important component of the electronic health record (EHR). As an increasing amount of CDS is implemented, it will be important that this be accomplished in a fashion that assists in clinical decision making without imposing unacceptable demands and burdens upon the provider's practice.

Objective: The objective of our study was to explore an approach that allows CDS to be clinician-friendly from a variety of perspectives, to build a prototype implementation that illustrates features of the approach, and to gain experience with a pilot framework for assessment.

Methods: The paper first discusses the project's design philosophy and goals. It then describes a prototype implementation (Neuropath/CDS) that explores the approach in the domain of neuropathic pain and in the context of the US Veterans Administration EHR. Finally, the paper discusses a framework for assessing the approach, illustrated by a pilot assessment of Neuropath/CDS.

Results: The paper describes the operation and technical design of Neuropath/CDS, as well as the results of the pilot assessment, which emphasize the four areas of focus, scope, content, and presentation.

Conclusions: The work to date has allowed us to explore various design and implementation issues relating to the approach illustrated in Neuropath/CDS, as well as the development and pilot application of a framework for assessment.

(JMIR Med Inform 2014;2(2):e20) doi:10.2196/medinform.3586

KEYWORDS

Internet; clinical decision support systems; electronic health records; neuropathic pain; therapeutics

Introduction

This paper describes a project that is exploring an approach to computer-based clinical decision support (CDS), built using Web technologies and linked to the electronic health record (EHR). The goal of the project is to explore ways in which CDS can be made clinician-friendly from a variety of perspectives. We believe that this is an important goal that needs to be addressed explicitly. This project is one step in that direction. The paper first discusses the project's design philosophy. It then describes a prototype implementation (Neuropath/CDS) that explores the approach in the domain of neuropathic pain, in the context of the US Department of Veterans Affairs (VA) EHR. Finally the paper discusses a framework for assessing the approach, illustrated by a pilot assessment of the prototype.

Computer-based CDS is common in health care systems, especially those that have an EHR [1]. CDS applications serve a variety of purposes, including providing alerts and reminders to clinicians at the point of care and making patient-specific recommendations about treatment and medications. Other approaches to computer-based CDS include computer-based clinical practice guidelines [2], clinical "dashboards" that provide focused information about specific clinical issues [3], and clinical "info buttons" [4] designed to facilitate access to online information relevant to a patient's clinical status.

The VA's national EHR (which includes the Veterans Health Information Systems and Technology Architecture backend and the Computerized Patient Record System (CPRS) interface) has evolved over the course of several decades as a powerful tool for patient care. There is widespread agreement that the CPRS can be made more powerful by incorporating increasing amounts of CDS. It will be important that this be accomplished in a fashion that assists the provider in clinical decision making without imposing additional demands and burdens upon the provider's practice. The present project complements research that is underway at other VA locations including the ATHENA system that provides decision support for hypertension [5], and for opioid therapy for chronic, noncancer pain [6,7].

Studies of CDS deployment have documented the importance of fitting CDS into the clinician's workflow, and of providing information at the time and place of clinical decision making [8]. The present work explores one approach to addressing these issues and a number of challenges that must be confronted, with a particular focus on accomplishing the goal in a clinically friendly fashion.

Methods

Design Philosophy

The overall goal of the project is to develop and explore an approach to computer-based CDS that is clinically efficient and clinician friendly. We want to include information that will be particularly useful to help the clinician manage a particular patient and make it easily accessible.

- The system should focus on a well defined, constrained clinical domain whose scope is easily understood by a clinician. Particularly important in this regard are the issues of focus and scope discussed below in the section describing our pilot framework for assessment.
- The system should be easy to use. For example: (1) it should be easy to invoke from the EHR, (2) it should be able to extract clinical data from the EHR rapidly, (3) the clinician should be able to inspect the information provided rapidly and easily, and (4) use of the system should be optional, not required.
- The system should provide a range of potentially useful information, accessible "in one place".
- The system should present material in a flexible intuitive fashion.
- The system should provide information to help the clinician decide how to manage the patient, but not try to tell the clinician what to do next.

A Prototype Implementation- Neuropath/CDS

Neuropath/CDS is a prototype CDS system developed to explore the issues involved in bringing computer-based CDS to the clinician at the point of care in a fashion that fits efficiently into the busy clinical environment. Neuropath/CDS uses patient data automatically extracted from the VA EHR to assist the primary care provider (PCP) in decisions regarding the first-line pharmacologic management of neuropathic pain (NP). NP is a chronic neurological condition that often requires continual monitoring and adjustment of treatment [9-11]. Ongoing clinical trials have generated repeated revisions of guidelines [9,12]. Multiple attempts at treatment with different drugs or drug combinations are often necessary, with drug effectiveness and side effects varying among different patients.

Neuropath/CDS has been refined in collaboration with the Neurology Service at the VA Connecticut Healthcare System (VACHS). Key features of the approach include the following.

- The system is an optional adjunct to care, accessible via a drop-down "Tools" menu from the EHR interface.
- It is driven from the patient record, retrieving patient demographics, comorbidity information, as well as current and past drug prescription information in a few seconds.
- The information provided is accessible from a single screen that we expect can be typically processed by the clinician user in well under a minute.
- Comments and recommendations are not prescriptive (ie, do not tell the clinician how to manage the patient), but rather describe options for management tailored to the patient's comorbidities and current treatment regimen.
- A visual outline of pharmacologic management is also provided, with "hover boxes" that provide further information about pharmacologic options. Several links to additional information are also provided.

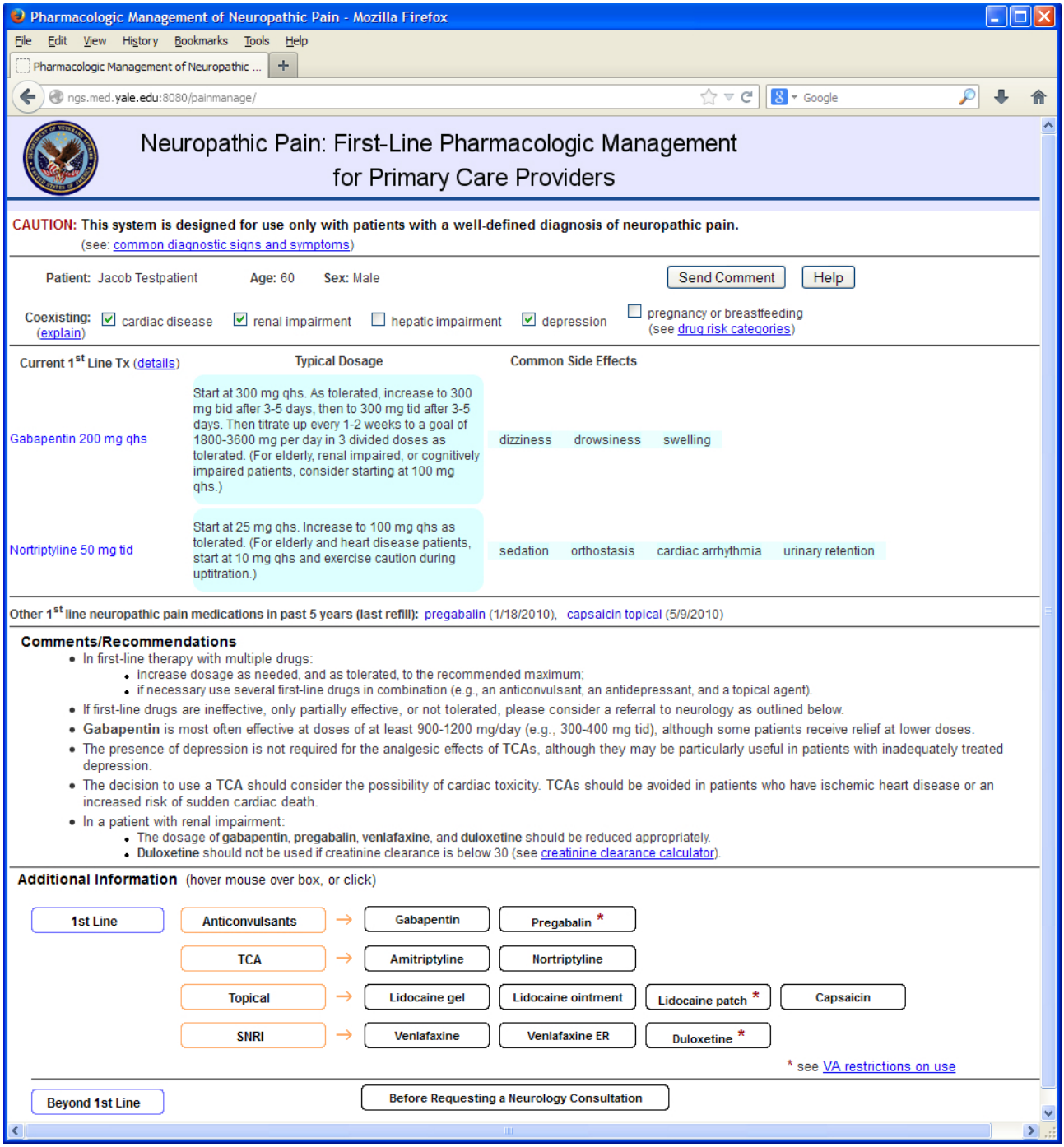
In this section, we first show an example of Neuropath/CDS in operation and describe the various components of its clinical user interface. We then describe the system's technical design, and our pilot framework for assessment.

Neuropath/CDS in Operation

Clinical Interface

Figure 1 shows the Web-based clinical interface of Neuropath/CDS. To view this interface, the clinician must select a patient within CPRS and invoke Neuropath/CDS using the "Tools" menu located at the top of the CPRS screen. The system typically takes 5-10 seconds to gather data from the EHR and present this interface. Since the system is still under development, a password is currently required. The clinical interface includes the following components.

Figure 1. The Web-based clinical interface of Neuropath/CDS, as described in the text. TCA=tricyclic antidepressant, SNRI=serotonin–norepinephrine reuptake inhibitor.



Basic Patient Data

Near the top of the screen are the patient's name (in this example, a test patient), age, sex, and a set of check boxes indicating the presence or absence of comorbid disease relevant to NP. These checkboxes are set automatically based on International Classification of Diseases, 9th Revision (ICD9) codes in the EHR, but can be changed (checked or unchecked) by the provider based on his/her knowledge of the patient's clinical status.

Current and Past Neuropathic Pain Medications

A little lower on the screen is information about the patient's current NP medications, which is followed by information about

any other NP medications taken during the past five years, all extracted automatically from the EHR. For the patient's current first-line NP medications, information about dosage and common side effects is also shown.

Comments and Recommendations

The next section shows a set of “Comments/Recommendations”. These items are generated by if-then logic, and are tailored to the patient’s age, sex, comorbid diseases, and current NP medications. The goal is not to try to tell the clinician what to do next, but rather to present relevant patient-specific issues to be considered in making such a decision.

A Visual "Hover Box" Outline of the Domain

The bottom of the screen presents a visual outline of first-line pharmacologic management of NP. If the user moves the computer mouse over one of the item names, a temporary “hover box” appears on the screen containing information about that item. Figures 2-4 show these boxes. The goal is to allow the clinician to explore alternatives for the patient's first-line NP management. The information presented can be conditionally tailored to the patient by if-then logic. If the user clicks on one of the item names, this information is presented in a pop-up box, which can be saved, printed, or copied into the patient record.

Figure 2. An example "hover box" that appears when the user "hovers" the mouse over the box labeled "Gabapentin" near the bottom of the screen shown in Figure 1.

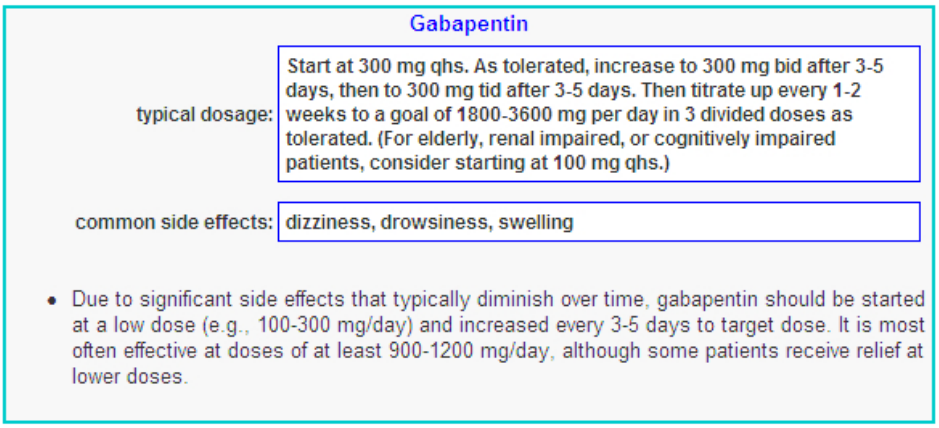


Figure 3. An example "hover box" linked to the box labeled "SNRI". SNRI=serotonin–norepinephrine reuptake inhibitor.

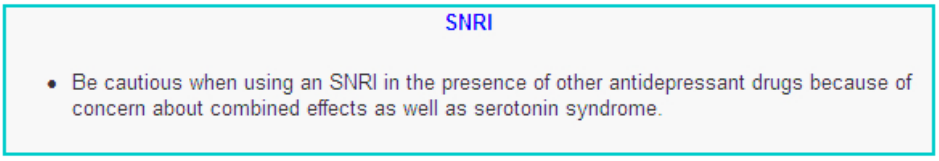
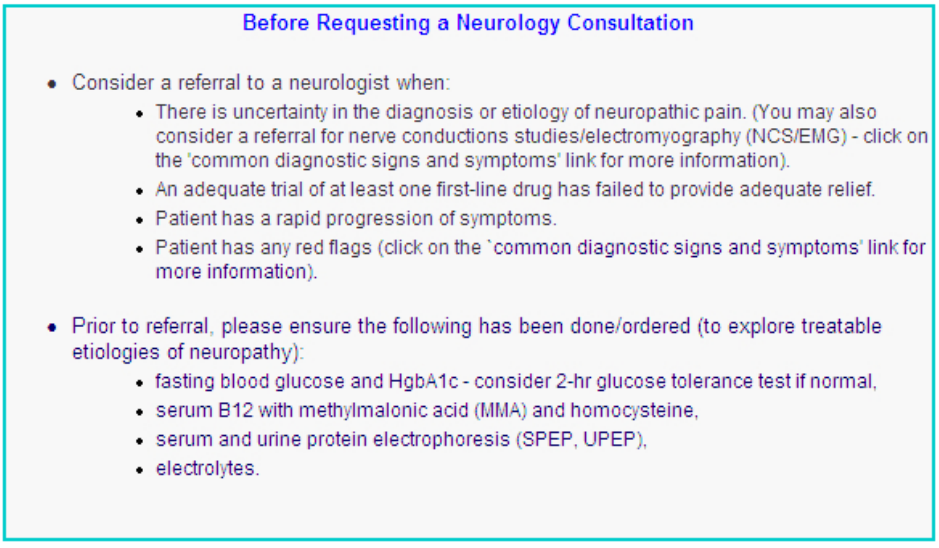


Figure 4. An example "hover box" linked to the box labeled "Before Requesting a Neurology Consultation".



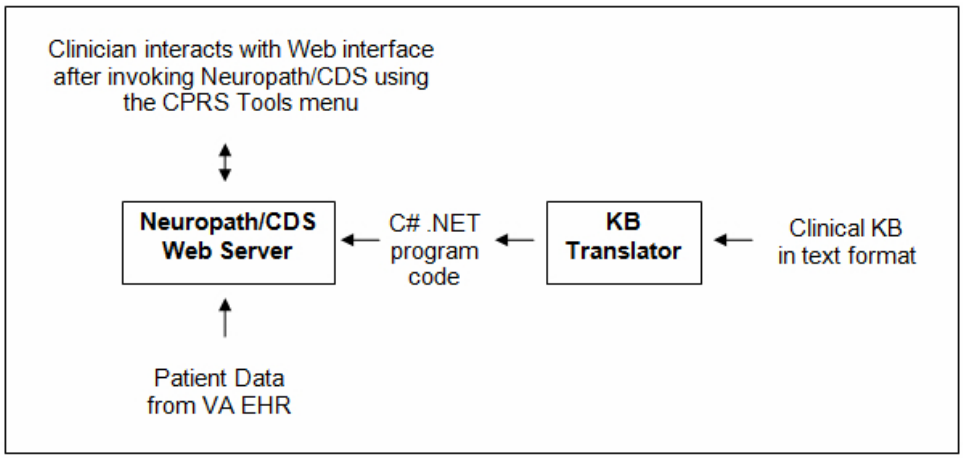
Links to Static Web Pages Providing Focused Clinical Information

The interface also provides a limited number of links to static Web pages that provide additional information, for instance: (1) to assist in making the diagnosis of NP, (2) to provide pregnancy and lactation risk information for NP medications, and (3) to provide instructions regarding the steps required before using drugs that are not part of the normal VA formulary.

Overview of Neuropath/CDS's Technical Design

Figure 5 shows a simplified schematic overview of Neuropath/CDS's technical design. When the clinician invokes Neuropath/CDS from CPRS using the Tools menu, a Web server program is activated with the patient's numeric identifier (ID) passed as a parameter. The Web server then retrieves patient data from the EHR as described below, and uses it to fill in the clinical values in the Web interface.

Figure 5. A simplified schematic outline of Neuropath/CDS's technical design. KB=knowledge base; CPRS=computerized patient record system; VA=Veterans Administration; EHR=electronic health record.



Automatic Extraction of Patient Data From the Electronic Health Record

The key factor in allowing Neuropath/CDS to operate with clinical data retrieved from the EHR is that this must be done quickly, in a clinically acceptable timeframe. Performing straightforward database calls to the EHR (which contains millions of rows of data not optimized for direct querying) would take much too long. We have taken two approaches to deal with this problem.

A variety of Web service procedures have been developed within VA EHR to facilitate the very rapid retrieval of selected patient-specific data. Neuropath/CDS uses these to retrieve the patient's current medications, from which the current NP medications are extracted.

For certain data (ICD9 codes and previous NP medications), a program runs through the entire VACHS patient database during the night to extract and condense the appropriate patient data into a much smaller "data staging" database that can be rapidly queried. For example, based on an examination of ICD9 codes in the EHR, a bit is set in the staging database reflecting the presence or absence of heart disease for each patient. Neuropath/CDS queries this smaller database to retrieve patient-specific data on comorbid disease and previous NP medications.

Clinical Knowledge Base

The Neuropath/CDS knowledge base (KB) consists of if-then logic of two types: (1) if-then rules, and (2) conditional comments.

The simple if-then rule shown in Textbox 1 states, "if the patient is receiving amitriptyline treatment or nortriptyline treatment, then the patient is receiving TCA (tricyclic antidepressant) treatment". Before the KB logic is executed, the variables amitriptyline_tx and nortriptyline_tx are set appropriately to true or false (by program code) based on a patient's clinical data.

The conditional comment shown in Textbox 2 states that: (1) "if the patient is receiving TCA treatment and has a history of depression", then a specified text comment should be included in the "Comments/Recommendations" section of the screen, and (2) "if the patient has a history of depression", then the text comment should also be included in the "TCA" hover box. In this way, text comments are directed to the different sections of the interface, where they are then sorted based on the "order" number (shown below), and displayed in sorted sequence.

Notice that this design allows a comment to be presented in different locations within the Web interface (eg, in the "Comments/Recommendations" section and/or in a "hover box"). A single comment can be directed to more than one screen location based on different if-then logic for each location. Comments can also be sorted and displayed in a different order in different locations.

The KB Translator automatically converts the KB into C# .NET code, which is then copied into a larger Web server program (built using Asynchronous JavaScript and Extensible Markup Language [AJAX], C# .NET, and cascading stylesheets) that implements Neuropath/CDS as a whole. By allowing the logic to be translated automatically into C# .NET code that is then

run directly on the Web server (compared, for example, to using an interpretive rules engine package or some other complex knowledge manipulation environment), we achieve logic that runs very rapidly. Using this approach, the Web interface is dynamic. For example, if the user checks or unchecks one the checkboxes indicating the presence/absence of comorbid disease, the KB logic is immediately rerun and the system's comments/recommendations on the Web screen are changed appropriately in a fraction of a second.

Textbox 1. A simple if-then rule.

```
if (amitriptyline_tx
    || nortriptyline_tx)        tca_tx = true;
```

Textbox 2. A conditional comment.

```
Comment TCA_depression {
Condition: tca_tx & depression_hx; Where: Recommendations (order: 7);
Condition: depression_hx; Where: TCA_comments (order: 5);
Text:        "The presence of depression is not required for the analgesic effects
              of TCAs, although they may be particularly useful in patients with
              inadequately treated depression." }
```

Developing a Framework for Assessment- Scope, Focus, Content, and Presentation

To provide a structure for our analysis, we developed a framework structured around the four areas of focus, scope, content, and presentation, as described below. This framework may also prove useful to CDS researchers or developers who might wish to adopt this general approach to CDS or certain of its features. Although the dividing line between these four areas is not always distinct, we believe that they provide a framework that is useful for discussing our experience and the lessons learned to date, and potentially useful to others in the future.

The current implementation of Neuropath/CDS is the result of several years of iterative development that has involved multiple interactions with many VA clinicians and technical staff. This process started as a collaboration with the VACHS Pain Management Clinic. As the project matured, the primary focus of the clinical collaboration shifted to the VACHS Neurology Service, and its relationship to VACHS PCPs, as described in more detail below.

In addition to extensive collaboration with individual clinicians, the system was presented at several conferences within the VACHS and elsewhere. These sessions resulted in further refinement of the system's design. We then conducted a formative study, with IRB approval, involving eight structured sessions in which project staff sat with a VACHS clinician as he/she used the operational CPRS interface with real patients and some test patients. During these sessions, the clinicians interacted directly with the computer using the mouse and keyboard, with project staff verbally providing a discussion of the features and answers to questions. The clinicians were asked to "think aloud" as they interacted with the various features of the system, vocalizing whatever thoughts they were having, which included reactions, comments, suggestions, and questions. At the end of the session, the clinician was asked to talk more broadly about the system, its features, its potential utility, as well as suggestions of additional features, content, and functionality.

We were interested in clinician reactions to the current Neuropath/CDS system, as well as to the overall approach to CDS. Although clinicians were very positive about the system and approach, we did not focus on trying to quantitate this aspect of their assessment, since they knew they were speaking directly to members of the system development staff, and evaluative comments would therefore be potentially biased. Instead, we focused on the other aspects of their comments.

Results

Developing and Refining the Approach

Since this paper describes a research project exploring an approach to providing CDS in a fashion that fits naturally into the clinical environment, the methodology described above is a major part of the "results" of the project. In addition, in this section we discuss the results of applying the pilot framework for the assessment described above.

Focus

The key factor that has shaped the refinement of Neuropath/CDS to its current design was the decision to focus the system's role on the clinical interface between the VACHS Neurology Service and VACHS PCPs. PCPs treat NP themselves, but when they need assistance they typically consult Neurology. Neuropath/CDS is therefore designed to help the PCP manage NP in the manner that Neurology would recommend. The system also suggests situations in which the PCP might decide to consult Neurology, and a recommended workup for the PCP to perform when requesting such a consult. Once defined, this focus on the interface between Neurology and PCPs provided a guiding context for all other aspects of the system's design.

Specific Comments on Focus

None of our VACHS clinician subjects commented explicitly on this focused role for the system. They appeared to find it natural and logical in the VACHS clinical environment.

Scope

Once this focus was defined, it became logical that the scope of the system's clinical domain should center on the first-line pharmacologic management of NP. This scope represents a markedly reduced subset of the possible clinical issues involving NP that we might have included. For example it does not include: (1) the diagnosis of NP, (2) second-line and third-line pharmacologic management of NP which involve the use of opioids and tramadol (second line) and drugs like carbamazepine, lamotrigine, topiramate, valproic acid, bupropion, citalopram, or paroxetine (third line), or (3) interventional nonpharmacologic approaches such as spinal cord stimulation, intravenous infusions, epidural injections, and nerve blocks.

The major reason for this restriction in scope was that the VACHS neurologists wanted to be consulted before PCPs went beyond first-line NP drugs. This also reflected a national VA goal to avoid the use of opioids as much as possible, and to have PCPs get as much utility out of first-line drugs as possible.

Specific Comments on Scope

An unexpected finding when clinicians used the system with real patients was that there was potential ambiguity regarding the borderline between first-line NP treatment and more advanced NP treatment, especially if one just looks at the drugs the patient is receiving. Some patients receiving first-line treatment for NP were also at the same time receiving an opioid (a second-line NP drug) for some other pain problem (not NP), or a high potency antidepressant (for clinically severe depression) that could also be used as a third-line NP drug. This was a potential source of confusion for the clinician in understanding the system's role and interpreting its advice for such patients.

To attempt to clarify these issues, we added additional explanatory text to the interface, adjusted the way in which the different NP medications were presented, and added the following comment to be used with patients who were receiving second- and third-line NP drugs.

> *This system is designed ONLY for first-line pharmacologic management of neuropathic pain (NP). Some patients receive second- or third-line NP drugs for other reasons (not for NP), in which case the comments below may still be relevant. If this patient is receiving a second-line drug (eg, an opioid or tramadol) or a third-line drug (eg, carbamazepine, lamotrigine, topiramate, valproic acid, bupropion, citalopram, or paroxetine) for NP, and you have concerns about NP management, please consider referral to neurology.* [Neuropath/CDS explanatory text]

As we gain more experience with the use of Neuropath/CDS in the clinical environment, we may need to make further refinements to deal with this particular issue involving scope.

An additional issue related to scope, which generated considerable discussion during the system's development, involved how to deal with the diagnosis of NP. The concern was how best to make sure that the system was used for appropriate patients (ie, patients who indeed had NP). An approach might have been to include what would essentially be a second complementary interactive CDS system focused on NP diagnosis. It was decided that this was not necessary and that this issue could be addressed by including a link to a static Web page that outlined in tabular form the clinical signs and symptoms that could serve to confirm the diagnosis of NP, or to suggest other diagnoses. All of our subject clinicians read this page quite carefully, and several explicitly stated that the content was very good. They did not appear to believe that a different approach was needed, and none expressed any concern about this approach to diagnosis of NP.

Content

Our main goal related to content was to provide useful practical information, relevant to immediate patient management. During the development of Neuropath/CDS, issues raised included the following.

A number of clinicians indicated that detailed dosage information would be very useful.

There was wide agreement that providing a list of first-line NP drugs that had been prescribed in the past would be particularly useful, since it could take up to 30-60 minutes of frustrating search through the EHR to find this information.

A succinct description of the steps required before a PCP could use a "nonformulary" drug was also identified as valuable. The VA document provided for each such drug is typically 6-8 pages long. The clinically relevant information could be condensed to a paragraph or two expressed as a bulleted outline.

A readily available outline of the pregnancy and lactation risk levels for the various NP drugs was also identified as useful to have available.

Many features such as these were added incrementally based on provider feedback as we developed and refined the system over time. Another major content-related goal was that the comments and recommendations made should not be prescriptive. The goal is to present relevant issues to be considered in making a decision in the context of a particular patient's current clinical status.

Specific Comments on Content

Different clinicians raised a number of specific issues. For example: (1) Might the system provide more assistance in choosing among first-line drugs? (2) Might the system more explicitly address the desirability of not using opioids prematurely? (3) Should the system be more explicit about the need to use lower doses of drugs in the elderly? (4) Is it really necessary to order serum and urine protein electrophoresis (tests to rule out multiple myeloma) when requesting a Neurology

consult? And (5) should the system recommend against the use of amitriptyline in a patient over 65 (a suggestion made by a geriatrician based on geriatric guidelines)?

The answers to questions of this sort are not always clear-cut. For example: (1) there is a great deal of latitude for practice variation and preference in the use of these drugs, and (2) VACHS PCPs are very familiar with the need to reduce dosage in the elderly, so stating this might just "clutter" the interface unnecessarily. The real lesson learned here is that there will always be room for judgment and iterative refinement in adjusting the content of the system's knowledge, based on user comments and domain expert judgment.

Presentation

From the standpoint of presentation, we wanted the system to have one page with a very limited number of links so that the PCP could easily preserve context when using the system. The use of hover boxes was designed to help in that regard. We also wanted the presentation to be as clear, intuitive, and helpful as possible.

Specific Comments on Presentation

Different clinicians raised a number of specific issues. It became clear that some features of the interface that seemed obvious to the developers were not immediately obvious to some of the clinicians using the system, for example: (1) the fact that the comorbidity checkboxes were initially set based on material from the chart (ICD9 codes), but could be changed by the PCP; or (2) that in the visual display of treatment options, the different treatment modalities were arranged in horizontal rows.

A clinician commented that there is quite a bit of material on this single screen. Another clinician felt that the use of color could enhance the readability of the static Web pages. Several clinicians were particularly positive about the ability to integrate use of the CDS with the CPRS interface, for example, the ease of access via the Tools menu, and the ability to copy and paste material from the CDS into the EHR.

These comments made it clear that it was important that a clinician should be shown the system, either in a clinical conference or by a colleague before using it, so he/she would be comfortable with its organization and understand the full range of its features.

Discussion

Current Status and Future Directions

Neuropath/CDS is currently operational on a pilot basis in the VACHS EHR, requiring a password since it is still under development. A logical future goal would be to make the approach available as adjunct to care with no password required. This would allow us to explore issues such as how best to promote its use, how frequently it is used, the types of patients it is used for, and the types of knowledge that tend to be accessed by PCPs (eg, which links and hover boxes are viewed).

We are also interested in extending this model of CDS, or components of the model, to other clinical domains. A number of clinicians suggested other potential domains for applying this model of CDS, including other types of pain, such as low back pain and headache, as well as many other diseases beyond pain. Some clinical domains may be sufficiently circumscribed that an approach very similar to that of Neuropath/CDS would work well. At the same time, it is clear that other domains may be much more complex and may require considerably more than a single screen to deal adequately with the clinical issues involved. For example, in the treatment of headache, there are at least five common types of headache, each with distinct approaches to pharmacologic management, and each with more first-line drugs than NP. In addition, one might like to provide some interactive assistance with diagnosis of headache. As a result, a CDS for a domain like headache might use features of Neuropath/CDS, but might well need to be considerably more complex.

Conclusions

A major source of CDS within the VA EHR currently consists of clinical alerts and reminders, which can be very time-consuming and frustrating for the PCP to manage. In building Neuropath/CDS, we wanted to provide a tool that is perceived as helpful and not burdensome to the busy PCP. Ideally the clinician could spend less than a minute interacting with the system, be exposed to information useful in managing a specific patient, and then move rapidly on to his/her next task.

Neuropath/CDS has been built and deployed on a prototype basis in the context of the VACHS EHR. The work to date has allowed us to explore: (1) various design and implementation issues relating to the system itself and to the underlying approach to CDS; as well as (2) a framework for assessment, with a particular emphasis on issues relating to focus, scope, content, and presentation.

Acknowledgments

This research was supported in part by National Institutes of Health grants UL1 RR024139 and T15 LM07056, VA grant REA 08-266, the VA Advanced Fellowship Program in Medical Informatics, and the VA Region 1 Specialty Care Access Network-Extension for Community Health care Outcomes. The authors would particularly like to thank Dr Gerald Grass, former Director of the VACHS Pain Management Clinic, who provided a great deal of assistance and support during the early stages of this project. The authors would also like to thank the numerous VACHS clinicians and researchers who provided comments and advice to help in the development and refinement of Neuropath/CDS. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the US Department of Veterans Affairs. The Human Investigations Committees at Yale University and at the VA Connecticut Healthcare System approved the study.

Conflicts of Interest

None declared.

References

1. Berner ES. AHRQ Publication No. 09-0069-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2009 Jun. Clinical decision support systems: State of the art. URL: http://healthit.ahrq.gov/sites/default/files/docs/page/09-0069-EF_1.pdf [accessed 2014-07-23] [WebCite Cache ID 6RHe0XBqd]
2. Shiffman RN, Michel G, Essaihi A, Thornquist E. Bridging the guideline implementation gap: A systematic, document-centered approach to guideline implementation. J Am Med Inform Assoc 2004;11(5):418-426 [FREE Full text] [doi: 10.1197/jamia.M1444] [Medline: 15187061]
3. Koopman RJ, Kochendorfer KM, Moore JL, Mehr DR, Wakefield DS, Yadamsuren B, et al. A diabetes dashboard and physician efficiency and accuracy in accessing data needed for high-quality diabetes care. Ann Fam Med 2011;9(5):398-405 [FREE Full text] [doi: 10.1370/afm.1286] [Medline: 21911758]
4. Del Fiol G, Curtis C, Cimino JJ, Iskander A, Kalluri AS, Jing X, et al. Disseminating context-specific access to online knowledge resources within electronic health record systems. Stud Health Technol Inform 2013;192:672-676 [FREE Full text] [Medline: 23920641]
5. Goldstein MK. Using health information technology to improve hypertension management. Curr Hypertens Rep 2008 Jun;10(3):201-207. [Medline: 18765090]
6. Trafton JA, Martins SB, Michel MC, Wang D, Tu SW, Clark DJ, et al. Designing an automated clinical decision support system to match clinical practice guidelines for opioid therapy for chronic pain. Implement Sci 2010;5:26 [FREE Full text] [doi: 10.1186/1748-5908-5-26] [Medline: 20385018]
7. Trafton J, Martins S, Michel M, Lewis E, Wang D, Combs A, et al. Evaluation of the acceptability and usability of a decision support system to encourage safe and effective use of opioid therapy for chronic, noncancer pain by primary care providers. Pain Med 2010 Apr;11(4):575-585. [doi: 10.1111/j.1526-4637.2010.00818.x] [Medline: 20202142]
8. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. BMJ 2005 Apr 2;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
9. Dworkin RH, O'Connor AB, Backonja M, Farrar JT, Finnerup NB, Jensen TS, et al. Pharmacologic management of neuropathic pain: Evidence-based recommendations. Pain 2007 Dec 5;132(3):237-251. [doi: 10.1016/j.pain.2007.08.033] [Medline: 17920770]
10. O'Connor AB, Dworkin RH. Treatment of neuropathic pain: An overview of recent guidelines. Am J Med 2009 Oct;122(10 Suppl):S22-S32. [doi: 10.1016/j.amjmed.2009.04.007] [Medline: 19801049]
11. Dworkin RH, O'Connor AB, Audette J, Baron R, Gourlay GK, Haanpää ML, et al. Recommendations for the pharmacological management of neuropathic pain: An overview and literature update. Mayo Clin Proc 2010 Mar;85(3 Suppl):S3-14 [FREE Full text] [doi: 10.4065/mcp.2009.0649] [Medline: 20194146]
12. Finnerup NB, Sindrup SH, Jensen TS. The evidence for pharmacological treatment of neuropathic pain. Pain 2010 Sep;150(3):573-581. [doi: 10.1016/j.pain.2010.06.019] [Medline: 20705215]

Abbreviations

CDS: clinical decision support
CPRS: computerized patient record system
EHR: electronic health record
ICD9: International Classification of Diseases, 9th Revision
KB: knowledge base
NP: neuropathic pain
PCP: primary care provider
TCA: tricyclic antidepressant
VA: Veterans Administration
VACHS: VA Connecticut Healthcare System

Edited by G Eysenbach; submitted 07.06.14; peer-reviewed by A Installe, C Schäfer; comments to author 30.06.14; revised version received 06.07.14; accepted 08.07.14; published 18.08.14.

Please cite as:
Miller P, Phipps M, Chatterjee S, Rajeevan N, Levin F, Frawley S, Tokuno H
Exploring a Clinically Friendly Web-Based Approach to Clinical Decision Support Linked to the Electronic Health Record: Design Philosophy, Prototype Implementation, and Framework for Assessment
JMIR Med Inform 2014;2(2):e20
URL: http://medinform.jmir.org/2014/2/e20/
doi:10.2196/medinform.3586
PMID:25580426

Original Paper

Clinical Decision Support System to Enhance Quality Control of Spirometry Using Information and Communication Technologies

Felip Burgos[1*], RN, MSc, PhD; Umberto Melia[2*], MBiomedEng; Montserrat Vallverdú[2*], PhDEng; Filip Velickovski[3,4*], B.BioMedSci(Hons), MSc; Magí Lluch-Ariet[3,5*], BSc., CSci, MSc; Pere Caminal[2*], PhDEng; Josep Roca[1*], MD, PhD

[1]Hospital Clinic - IDIBAPS - Ciberes, Respiratory Diagnostic Center, University of Barcelona, Barcelona, Spain

[2]Centre de Recerca en Enginyeria Biomèdica (CREB-UPC), Universitat Politècnica de Catalunya, Barcelona, Spain

[3]Barcelona Digital Technology Centre, Barcelona, Spain

[4]ViCOROB, Universitat de Girona, Girona, Spain

[5]Departament d'Enginyeria Telemàtica (ENTEL), Universitat Politècnica de Catalunya, Barcelona, Spain

[*]all authors contributed equally

Corresponding Author:
Felip Burgos, RN, MSc, PhD
Hospital Clinic - IDIBAPS - Ciberes
Respiratory Diagnostic Center
University of Barcelona
Sotano porta 6
Villarroel, 170
Barcelona, 08036
Spain
Phone: 34 932275540
Fax: 34 932275455
Email: fburgos@ub.edu

Abstract

Background: We recently demonstrated that quality of spirometry in primary care could markedly improve with remote offline support from specialized professionals. It is hypothesized that implementation of automatic online assessment of quality of spirometry using information and communication technologies may significantly enhance the potential for extensive deployment of a high quality spirometry program in integrated care settings.

Objective: The objective of the study was to elaborate and validate a Clinical Decision Support System (CDSS) for automatic online quality assessment of spirometry.

Methods: The CDSS was done through a three step process including: (1) identification of optimal sampling frequency; (2) iterations to build-up an initial version using the 24 standard spirometry curves recommended by the American Thoracic Society; and (3) iterations to refine the CDSS using 270 curves from 90 patients. In each of these steps the results were checked against one expert. Finally, 778 spirometry curves from 291 patients were analyzed for validation purposes.

Results: The CDSS generated appropriate online classification and certification in 685/778 (88.1%) of spirometry testing, with 96% sensitivity and 95% specificity.

Conclusions: Consequently, only 93/778 (11.9%) of spirometry testing required offline remote classification by an expert, indicating a potential positive role of the CDSS in the deployment of a high quality spirometry program in an integrated care setting.

(JMIR Med Inform 2014;2(2):e29) doi:10.2196/medinform.3179

KEYWORDS

spirometry; telemedicine; information communication technologies; primary care; quality control

Introduction

High Quality Spirometry Testing

High quality spirometry testing across health care levels is pivotal for proper management of patients with prevalent chronic respiratory disorders, namely asthma and chronic obstructive pulmonary disease (COPD) [1].

We have recently reported the effectiveness of a Web-based application for remote offline expert support to enhance the quality of spirometry in primary care. High quality testing improved in a sustainable manner with the remote support [2]. A relevant difference was observed between the intervention group, 2419/3383 (71.50%) high quality spirometry, and the control group, 713/1198 (59.52%) high quality spirometry, throughout the 12 month follow-up period (P<.001). Similar figures have been obtained in pharmacy offices, as part of a COPD case finding program [3].

In the Basque Country (Spain), the ongoing regional deployment of the Web-based offline support program from specialists to primary care will cover the entire population, 2.2 million inhabitants, by the end of 2014 [4,5]. Interestingly, their results [6] are similar to those reported in the initial randomized controlled trial [2] described above.

Ideally, extensive deployment of a high quality spirometry program in the community should offer accessibility to standardized raw spirometric data through a technological architecture providing interoperability across health care tiers. To this end, a Clinical Document Architecture for spirometry using Health Level Seven v3 standards was recently made available by the Catalan Health Department [7], such that spirometric testing will be available at the regional level. In this scenario, automatic assessment of quality of spirometry testing should enhance the efficiency of the program. Unfortunately, current applications for online assessment of quality of spirometry misclassify the tests, as compared with examinations done by expert professionals [2].

Clinical Decision Support System

We hypothesize that elaboration and validation of a clinical decision support system (CDSS) for online automatic assessment and certification of quality of spirometry in primary care may represent a pivotal step toward regional adoption of the high quality spirometry program with an integrated care approach.

The current study is part of the refinement of the ongoing deployment of the high quality spirometry program in Catalonia [8], an European region of 7.5 million inhabitants.

Methods

Building-Up the Clinical Decision Support System

Figure 1 shows the initial step in the process for elaboration of the CDSS was the identification of the optimal sampling frequency to achieve the highest sensitivity and specificity in the analysis of the spirometric curves. To this end, a systematic examination of a large range of sampling frequencies, from 6.25 Hz to 100 Hz, was done during the first iterative process.

The process was done using the 24 standard flow-volume and volume-time curves from the pulmonary waveform generator recommended by the American Thoracic Society/European Respiratory Society (ATS/ERS) [7]. This set of 24 standard curves cover the entire spectrum of clinical abnormalities, as well as common spirometric artifacts. They are used as a reference material for calibration purposes and, in general, to facilitate comparisons among lung function laboratories.

The construction of an initial version of the CDSS was carried out using the 24 standard spirometry curves [9,10] following an iterative process, as displayed in Figure 1. In each step, the results generated by the CDSS were compared with the criteria of one expert in the field of lung function testing (FB), and the iterative process was maintained until sensitivity and specificity of the results generated by the CDSS showed 24/24 (100.0%) agreement with the expert.

The CDSS combines the different aspects assessed on the spirometry curve in one score with three different categories: (1) grade 0, rejected due to unacceptable morphology of the spirometry curve; (2) grade 1, acceptable for further classification according to Table 1; or (3) grade 2, undefined characteristics of the spirometry (see Multimedia Appendix 1 for examples of the three categories in Figure 1S). The two first categories, grades 0 and 1, allow proper online automatic classification of spirometry testing as well as the generation of a certified spirometry curve to be potentially shared across health care tiers; whereas grade 2 requires offline expert assessment.

Table 1. Quality scores for spirometric maneuvers according to ATS/ERS standardization [9].

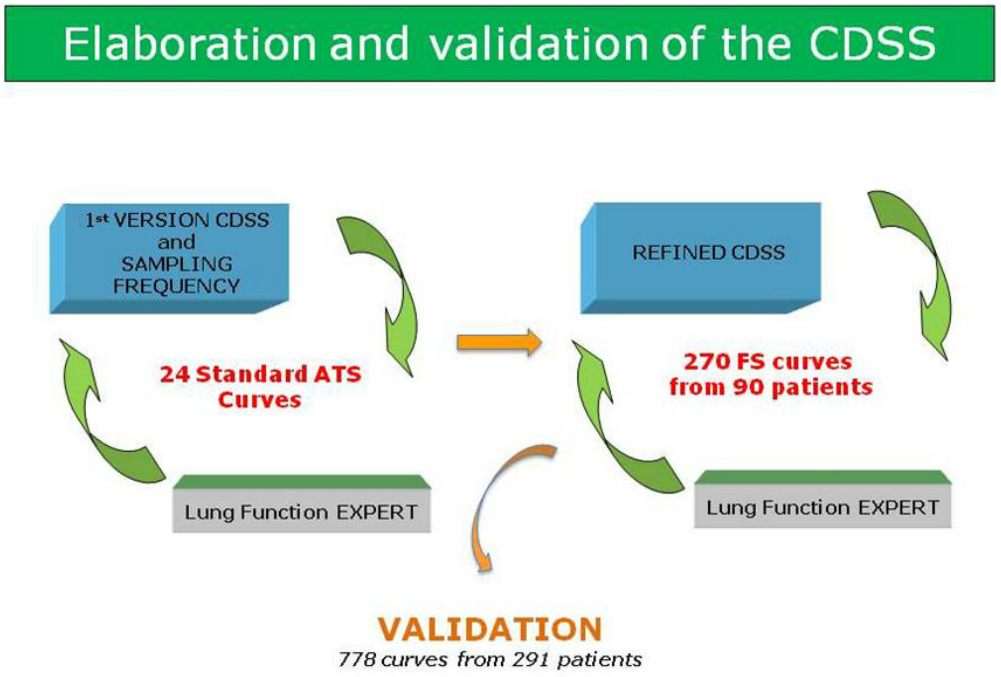
Scores	Maneuvers
A[a]	3 acceptable maneuvers, and best 2 matched with differences in FVC[b]and/or FEV_1<150 ml
B	3 acceptable maneuvers, and best 2 matched with differences in FVC[b]and/or FEV_1[c]<200 ml
C	2 acceptable maneuvers, and best 2 matched with differences in FVC and/or FEV_1[c]<250 ml
D	1 acceptable maneuver
F	No acceptable maneuvers

[a]High quality spirometries, A and B scores, correspond to A, 3 acceptable maneuvers with differences in FVC and/or FEV_1<150 ml; and B, 3 acceptable maneuvers with differences in FVC and/or FEV_1<200 ml; C, to high variability among maneuvers; D, only one acceptable maneuver; and F no acceptable maneuver.

[b]FVC = forced vital capacity

[c]FEV_1= forced expiratory volume in the first second

Figure 1. Flow of the process followed to elaborate and validate the Clinical Decision Support System (CDSS). ATS=American Thoracic Society; FS=forced spirometry.



The Characteristics and the Algorithm

The CDSS systematically assessed 27 different characteristics of each spirometry curve, as displayed in Table 2. There were four out of the 27 characteristics that were extracted from the international recommendations for standardization of the test, jointly reported by the ATS and the ERS [11]; whereas the remaining 23 were introduced during the current research. Each of these 27 features had a well defined specific algorithm for calculations. The mathematical description of a feature constituted the so-called metric. It is of note that a given feature may require more than one metric. The quantitative values of a given metric were denominated thresholds that were used for quality assessment. It is also of note that some metrics may have primary and secondary thresholds. The initial parameters of the automatic algorithm for online assessment of quality of spirometry were refined through successive iterations until the final version of the CDSS was obtained (Figure 1). As indicated above, the performance of each of the successive versions of the CDSS was compared with the results provided by the expert. A refined version of the CDSS was achieved using 270 curves from 90 patients from [2].

Table 2. List of criteria of the forced spirometry curve explored by the CDSS.

Forced spirometry curve	Criteria[i]
BEV[a] trad	Back extrapolation >0.15 L or < 5% of FVC[g]
EOTV[b] trad	End of test criteria, volume < 0.025 L in time ≥1 s
Tex[c]	Time of end FVC[g](Tex>6 s)
	a) EOTV[b]< 0.025 L or Tex[c]>6 s;
	b) If Tex[c]>6 s EOTV[b]<0.025 L in time 0.5 s;
	c) If Tex[c]>6 s, EOTV[b]< 0.1 L;
	d) EOTV[b](Tex[c]) < 0.025 L; and
EOTV[b] new (5 criteria)	e) EOTV[b]< 0.025 * Tex/6 L
Peak_Valley_Single	High local maximum (peak) and minimum (valley) in FV[e]curve
Peak_Valley_Combined	High local maximum (peak) and minimum (valley) in FV[e]curve close to FEV_1[h]
VT[d] end	Irregularity or oscillation at the end of FT[m]curve
FV[e]_slope_single	Variation of FV[e]slope or high FV[e]slope
FV[e]_slope_combined	Variation of FV[e]slope and high FV[e]slope
FV[e]Slope_Test_Combo	Irregularity and variation of FV[e]slope or high FV[e]slope
FV[e]Slope_Test_Combo_Area Under Rect[j]	Irregularity or variation of FV[e]slope and high FV[e]slope
FV[e]Slope_Test_Combo4	Irregularity and variation of FV[e]slope and high FV[e]slope
Diff_single[k]	Irregular concavity-convexity before the PEF[f]value in FV[e]curve
Diff_combined[l]	Irregular slope and irregular concavity-convexity before the PEF[f]value in FV[e]curve
PEF[f] TimeUp	Time to archive PEF[f]< 130 milliseconds
PEF[f] TimeDown	Time to archive PEF[f]> 0.25 milliseconds
PEF[f] Cut	PEF[f]is not a peak in FV[e]curve (is plane), volume (F^n=PEF[f]) > 15 % FVC[g]
PEF[f] Cut2 FEV_1[h]	PEF[f]is not a peak in FV[e]curve (is plane), volume (F^n=PEF[f]) > 17.5 % FEV_1[h]
PEF[f] DoublePeak	PEF[f]bimodal in FV[e]curve
PEF[f] Slow	Volume to archive PEF[f]< 20% FVC[g]

[a]BEV = back extrapolation

[b]EOTV = end of test criteria, volume

[c]Tex = Time to end FVC

[d]VT = volume/time curve

[e]FV = flow/volume curve

[f]PEF = peak expiratory flow

[g]FVC = forced vital capacity

[h]FEV_1= forced expiratory volume in the first second

[i]The list includes the classical parameters used by ATS/ERS guidelines [11].

[j]Rect = rectum

[k]Diff single= irregular concavity-convexity before the PEFf value in flow volumen curve concavity or convexity exists if the extracted signal metric

[l]Diff_combined = irregular slope and irregular concavity-convexity before the peak expiratory flow value in flow volume curve

[m]FT = flow/time curve

[n]F=flow

Clinical Decision Support System Validation

The refined version of the CDSS was compared with a database of 778 curves from 291 patients from one of the primary care centers in Barcelona. The spirometry testing was done using a spirometer (Sibel 120, SIBELMED, Barcelona Spain). Again, the score generated by the CDSS was compared with the one obtained from the same expert evaluator.

The use of the two patient databases, for refinement and validation purposes, was approved by the Ethical Committee of the Hospital Clínic i Provincial de Barcelona.

Data Analysis

The ATS database [10] contains volume (*V*) values of each curve, from which flow (*F*) values were obtained by discrete differentiation (equation 1, Figure 2). The two patient's databases contained *F* values, from which *V* values were obtained by discrete integration (equation 2, Figure 2). The sample period is Δt=0.01s, so the sample frequency is 100 Hz. Sensitivity and specificity of the CDSS were calculated for all curves classified as grades 0 or 1 using equations 3 and 4 in Figure 2.

Figure 2. Equations for data analysis. F=flow; V=volume; i=1,…,N; N=length of the sequence; true positive (TP) corresponds to curves classified as grade 0 by both CDSS and the evaluator; true negative (TN) corresponds to curves classified as grade 1 by the CDSS and the by the evaluator; false positive (FP) indicates curves classified as grade 0 by the CDSS, but classified in grade 1 by the evaluator; and, false negative (FN) corresponds to curves classified as grade 1 by the CDSS, but as grade 0 by the evaluator.

$$F[i]=V[i]-V[i-1] \,/\, \Delta t \quad (1)$$

$$V[i] = \sum_{n=0}^{i} F[i-n] * \Delta t \quad (2)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN+FP}, \quad (4)$$

Results

The Sampling Frequency

The sampling frequency that provided the highest sensitivity and specificity for the analysis carried out with the 24 standard spirometry curves recommended by the ATS [10] was 100 Hz (Figure 1 and Multimedia Appendix 2, Table 2S), this frequency is widely used in commercial spirometers, and it is reasonable from the electronic transferability point of view. This result was confirmed in the 270 curves from 90 subjects [2].

Both sensitivity and specificity of the CDSS were initially calculated with the 24 standard spirometry curves recommended by the ATS [11] using only grade 0 and grade 1 curves. The results were as follows, grade 0, n=15; grade 1, n=6; grade 2, n=3 with 24/24 (100.0%) sensitivity and 24/24 (100.0%) specificity. Up to five complete versions of the CDSS were generated in the two iterative processes indicated in Figure 1, until a final version of the CDSS was ready for validation.

Grading the Curves

The validation study using 778 curves from 291 patients showed the following distribution of spirometry curves, 419/778 maneuvers (53.8%) were appropriately classified as bad curves (grade 0); 266/778 maneuvers (34.2%) were appropriately classified as good curves (grade 1); and only 93/778 maneuvers (11.9%) needed an offline review by a lung function expert to assess quality of the test (grade 2; see Multimedia Appendix 3). Sensitivity and specificity calculations for grade 0 and grade 1 curves were 96.1 and 94.9%, respectively.

Discussion

The Current Research

The current research has generated and validated a CDSS that shows the ability to classify a reasonable percentage of spirometry curves, 685/778 (88.1%) as either acceptable (grade 1) or bad maneuvers (grade 0). Only 93/778 (11.9%) of the curves were classified as undefined (grade 2) and were candidates for offline remote validation by an expert. Moreover, we observed that both sensitivity and specificity of the CDSS were very high. Consequently, the results seem to indicate that a vast majority of spirometry testing carried out by nonspecialized professionals in primary care can be reliably assessed online, and the high quality spirometry program partly based on remote automatic evaluation of the testing could be considered ready for regional scalability. Obviously, further steps toward extensive deployment of the program must be planned with caution. A proper monitoring of the potential for generalization of the current results and the need for further refinements of the current CDSS should be taken into account.

The results of the current research overcome some of the limitations of the existing computer-based algorithms generating automatic feedback, as reported in [2,12]. It is acknowledged, however, that automatic feedback based on enhanced algorithms like the one proposed by the current research may be effective only if they are part of a comprehensive program for high quality forced spirometry.

In the new scenario, as indicated by the business process management notation (BPMN) diagram (Multimedia Appendix 2, Figure 2S), acceptable maneuvers (grade 1) will be automatically addressed to the algorithm indicated in Table 1 that classifies and certifies spirometry testing prior to its recording into the local (electronic health record) and regional

repositories. In contrast, those maneuvers classified as bad curves (grade 0) will generate an online specific error message to the professional, indicating the need to perform additional testing until quality acceptance is reached. As indicated, we estimate that approximately 12% of the curves will not be properly classified (grade 2), and they will need an offline remote supervision by an expert professional. In this case, the spirometry testing of a given patient may need to be rescheduled.

Previous reports have indicated the potential of telemedicine to enhance both quality and diagnostic potential of spirometry testing carried out by nonexpert professionals [13-15], but the quality control in those studies was based on offline analyses by expert professionals carried out in a time consuming manner [16-18]. Likewise, the need for an external, likely centralized, quality control program [15,17-20] is well established. The results of the current study refine previous achievements [2] and open the way to explore extensive and efficient adoption of this type of high quality spirometry programs.

We acknowledge that high quality spirometry programs combine several different dimensions, namely: (1) professional coaching [21,22]; (2) remote support [2]; (3) interoperability of testing across health care levels [20]; (4) standards for procurement of equipment [11,23]; and (5) support to interpretation of testing [24,25]. The current study provides pivotal results to efficiently address issues associated to remote support of spirometry testing. But, a proper integration of all the above elements needs to be considered in the process of shaping a successful high quality spirometry program for scalability at regional level.

Limitations of the Study

We acknowledge two principal limitations of the study. First, we included only one expert observed (FB). The CDSS should be reassessed in the future with the inclusion of at least 3 different experts. Moreover, the current study evaluates the CDSS in an isolated manner. But, further assessment of the whole clinical process as defined in the BPMN (see Multimedia Appendix 2, Figure 2S) should be done before specific plans for scalability are undertaken.

Conclusions

To our knowledge, the current study constitutes the first successful attempt to validate an automatic CDSS for large scale online assessment of quality of spirometry testing. The incorporation of the CDSS into the Web-based application for remote assistance to primary care professionals [2] may facilitate sustainable high quality spirometry generating a significant added value in an integrated care scenario.

The results indicate a high potential of the CDSS for discrimination between good and poor quality results of spirometry testing, but they require further independent validation before specific plans for implementation are materialized.

Acknowledgments

The authors thank Jordi Giner of Hospital de la Santa Creu i Sant Pau, in Barcelona for providing the validation database. This project was supported by Inforegió (AGAUR) 2008; NEXES (Supporting Healthier and Independent Living for Chronic Patients and Elderly, CIP-ICT-PSP-2007-225025); FIS PI09/90634. Servicios Innovadores de Atencion Integrada para Pacientes Crónicos - PITES- ISCIII 2010-12; EC-FP7 Programme, Synergy-COPD, GA nº 270086; TAMESIS (TEC2011-22746, Spanish Government) CIBER of Bioengineering, Biomaterials and Nanomedicine; Research Fellowship Grant FPU AP2009-0858 from the Spanish Government; and, Catalan Master Plan of Respiratory Diseases (PDMAR).

Conflicts of Interest

None declared.

Multimedia Appendix 1

The algorithm for computing maneuver acceptability, using the 27 set of criteria.

[PDF File (Adobe PDF File), 6KB - medinform_v2i2e29_app1.pdf]

Multimedia Appendix 2

Three examples with curves classified as Grade 0, 1 and 2. The Business Process Model Notation (BPMN) diagram displays the use of the CDSS for quality control in primary care within a coordinated care scenario. The results of the protocol undertaken to identify the optimal sampling frequency during the first iterative process are shown here.

[PDF File (Adobe PDF File), 113KB - medinform_v2i2e29_app2.pdf]

Multimedia Appendix 3

For each FS curve, the results generated by the CDSS are compared with those provided by the expert professional. It is of note, that only the expiratory portion of the FS manouevres was taken into account for analysis.

[PDF File (Adobe PDF File), 27KB - medinform_v2i2e29_app3.pdf]

References

1. Celli BR. The importance of spirometry in COPD and asthma: Effect on approach to management. Chest 2000 Feb;117(2 Suppl):15S-19S. [Medline: 10673468]
2. Burgos F, Disdier C, de Santamaria EL, Galdiz B, Roger N, Rivera ML, e-Spir@p Group. Telemedicine enhances quality of forced spirometry in primary care. Eur Respir J 2012 Jun;39(6):1313-1318 [FREE Full text] [doi: 10.1183/09031936.00168010] [Medline: 22075488]
3. Castillo D, Burgos F, Giner J, Estrada M, Soriano JB, Flor X. Arch Bronconeumologia. 2011. Cribado de EPOC: nuevas herramientas y circuitos sanitarios. Resultados preliminares del Farmaepoc II. (FARMAEPOC) URL: http://www.researchgate.net/publication/266081643 [accessed 2014-10-03] [WebCite Cache ID 6T3e56yDp]
4. Eustat - population and housing census in Euskadi. 2011. URL: http://www.eustat.es/estadisticas/tema_159/opt_0/ti_Poblacion/temas.html [accessed 2014-09-11] [WebCite Cache ID 6SW3bUQcx]
5. Marina Malanda N. Agencia de Evaluación de Tecnologías Sanitarias del País Vasco. Informes de evaluación de tecnologías sanitarias: OSTEBA Nº; 2012. URL: http://www.msssi.gob.es/organizacion/sns/planCalidadSNS/ [accessed 2014-09-25] [WebCite Cache ID 6Sr9RmROU]
6. Marina Malanda N, López de Santa María E, Gutiérrez A, Bayón J, Garcia L, Gáldiz J. Telemedicine spirometry training and quality assurance program in primary care centers of a public health system. Telemed J E Health 2014 Apr;20(4):388-392. [doi: 10.1089/tmj.2013.0111] [Medline: 24476193]
7. Salas T, Rubies C, Gallego C, Muñoz P, Burgos F, Escarrabill J. Technical requirements of spirometers in the strategy for guaranteeing the access to quality spirometry. Arch Bronconeumol 2011 Sep;47(9):466-469 [FREE Full text] [doi: 10.1016/j.arbres.2011.06.005] [Medline: 21821333]
8. Tresserras R, en nombre del Grupo de Trabajo de Planes Directores. Planning driven by health priorities. Master planning criteria. Med Clin (Barc) 2008 Dec;131 Suppl 4:42-46. [Medline: 19195477]
9. ATS. Am Rev Respir Dis. 1979 May. ATS statement--snowbird workshop on standardization of spirometry URL: http://www.unboundmedicine.com/medline/citation/453705/ATS_statement__Snowbird_workshop_on_standardization_of_spirometry_ [accessed 2014-09-25] [WebCite Cache ID 6Sr9wWlIm]
10. Am Rev Respir Dis. 1987 Nov. Standardization of spirometry: A summary of recommendations from the American Thoracic Society: The 1987 update URL: http://annals.org/article.aspx?articleid=701103 [accessed 2014-09-25] [WebCite Cache ID 6SrAEBfpy]
11. Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, ATS/ERS Task Force. Standardisation of spirometry. Eur Respir J 2005 Aug;26(2):319-338 [FREE Full text] [doi: 10.1183/09031936.05.00034805] [Medline: 16055882]
12. Müller-Brandes C, Krämer U, Gappa M, Seitner-Sorge G, Hüls A, von Berg A, et al. LUNOKID: Can numerical American Thoracic Society/European Respiratory Society quality criteria replace visual inspection of spirometry? Eur Respir J 2014 May;43(5):1347-1356. [doi: 10.1183/09031936.00058813] [Medline: 24232698]
13. Masa JF, González M, Pereira R, Mota M, Riesco JA, Corral J, et al. Validity of spirometry performed online. Eur Respir J 2011 Apr;37(4):911-918 [FREE Full text] [doi: 10.1183/09031936.00011510] [Medline: 20650985]
14. Bellia V, Pistelli R, Catalano F, Antonelli-Incalzi R, Grassi V, Melillo G, et al. Quality control of spirometry in the elderly. The SA.R.A. study. Salute respiration nell'Anziano = Respiratory health in the elderly. Am J Respir Crit Care Med 2000 Apr;161(4 Pt 1):1094-1100. [doi: 10.1164/ajrccm.161.4.9810093] [Medline: 10764296]
15. Spirometry monitoring technology - SPIROLA. Centers for Disease Control and Prevention. NIOSH - Spirometry in occupational medicine URL: http://www.cdc.gov/niosh/topics/spirometry/spirola-software.html [accessed 2014-10-02] [WebCite Cache ID 6T3eEkp9q]
16. Montes de Oca M, Tálamo C, Perez-Padilla R, Jardim JR, Muiño A, Lopez MV, PLATINO Team. Chronic obstructive pulmonary disease and body mass index in five Latin America cities: The platino study. Respir Med 2008 May;102(5):642-650 [FREE Full text] [doi: 10.1016/j.rmed.2007.12.025] [Medline: 18314321]
17. Pérez-Padilla R, Vázquez-García J, Márquez M, Menezes AM, PLATINO Group. Spirometry quality-control strategies in a multinational study of the prevalence of chronic obstructive pulmonary disease. Respir Care 2008 Aug;53(8):1019-1026 [FREE Full text] [Medline: 18655739]
18. Janssens W, Liu Y, Liu D, Kesten S, Tashkin DP, Celli BR, et al. Quality and reproducibility of spirometry in COPD patients in a randomized trial (UPLIFT®). Respir Med 2013 Sep;107(9):1409-1416. [doi: 10.1016/j.rmed.2013.04.015] [Medline: 23714653]
19. Enright PL. How to make sure your spirometry tests are of good quality. Respir Care 2003 Aug;48(8):773-776 [FREE Full text] [Medline: 12890297]
20. Enright PL, Skloot GS, Cox-Ganser JM, Udasin IG, Herbert R. Quality of spirometry performed by 13,599 participants in the World Trade Center Worker and Volunteer Medical Screening Program. Respir Care 2010 Mar;55(3):303-309 [FREE Full text] [Medline: 20196879]
21. Steenbruggen I, Mitchell S, Severin T, Palange P, Cooper BG, Spirometry HERMES Task Force. Harmonising spirometry education with HERMES: Training a new generation of qualified spirometry practitioners across Europe. Eur Respir J 2011 Mar;37(3):479-481 [FREE Full text] [doi: 10.1183/09031936.00187810] [Medline: 21357919]

22. Escarrabill J, Roger N, Burgos N, Giner J, Molins A, Tresserras R. Diseño de un programa de formación básico para conseguir espirometrías de calidad. Educ. méd 2012 Jun;15(2):103-107. [doi: 10.4321/S1575-18132012000200008]
23. Miller MR, Crapo R, Hankinson J, Brusasco V, Burgos F, Casaburi R, ATS/ERS Task Force. General considerations for lung function testing. Eur Respir J 2005 Jul;26(1):153-161 [FREE Full text] [doi: 10.1183/09031936.05.00034505] [Medline: 15994402]
24. Pellegrino R, Viegi G, Brusasco V, Crapo RO, Burgos F, Casaburi R, et al. Interpretative strategies for lung function tests. Eur Respir J 2005 Nov;26(5):948-968 [FREE Full text] [doi: 10.1183/09031936.05.00035205] [Medline: 16264058]
25. Pellegrino R, Brusasco V, Viegi G, Crapo RO, Burgos F, Casaburi R, et al. Definition of COPD: Based on evidence or opinion? Eur Respir J 2008 Mar;31(3):681-682 [FREE Full text] [doi: 10.1183/09031936.00154307] [Medline: 18310402]

Abbreviations

ATS: American Thoracic Society
BPMN: business process management notation
CDSS: clinical decision support system
COPD: chronic obstructive pulmonary disease
ERS: European Respiratory Society

Edited by G Eysenbach; submitted 06.04.14; peer-reviewed by P Enright, M Stanbrook; comments to author 16.06.14; revised version received 27.07.14; accepted 21.08.14; published 21.10.14.

Please cite as:
Burgos F, Melia U, Vallverdú M, Velickovski F, Lluch-Ariet M, Caminal P, Roca J
Clinical Decision Support System to Enhance Quality Control of Spirometry Using Information and Communication Technologies
JMIR Med Inform 2014;2(2):e29
URL: http://medinform.jmir.org/2014/2/e29/
doi:10.2196/medinform.3179
PMID:25600957

Original Paper

OWLing Clinical Data Repositories With the Ontology Web Language

Raimundo Lozano-Rubí[1,2], MD, BSCS; Xavier Pastor[1], MD, PhD; Esther Lozano[3], MSc

[1]Hospital Clínic, Unit of Medical Informatics, University of Barcelona, Barcelona, Spain

[2]Autonomous University of Barcelona, Department of Computer Science, Barcelona, Spain

[3]Universidad Politécnica de Madrid, Ontology Engineering Group, Madrid, Spain

Corresponding Author:
Raimundo Lozano-Rubí, MD, BSCS
Hospital Clínic
Unit of Medical Informatics
University of Barcelona
Villarroel 170
Barcelona, 08036
Spain
Phone: 34 932279206
Fax: 34 932279240
Email: rlozano@clinic.ub.es

Abstract

Background: The health sciences are based upon information. Clinical information is usually stored and managed by physicians with precarious tools, such as spreadsheets. The biomedical domain is more complex than other domains that have adopted information and communication technologies as pervasive business tools. Moreover, medicine continuously changes its corpus of knowledge because of new discoveries and the rearrangements in the relationships among concepts. This scenario makes it especially difficult to offer good tools to answer the professional needs of researchers and constitutes a barrier that needs innovation to discover useful solutions.

Objective: The objective was to design and implement a framework for the development of clinical data repositories, capable of facing the continuous change in the biomedicine domain and minimizing the technical knowledge required from final users.

Methods: We combined knowledge management tools and methodologies with relational technology. We present an ontology-based approach that is flexible and efficient for dealing with complexity and change, integrated with a solid relational storage and a Web graphical user interface.

Results: Onto Clinical Research Forms (OntoCRF) is a framework for the definition, modeling, and instantiation of data repositories. It does not need any database design or programming. All required information to define a new project is explicitly stated in ontologies. Moreover, the user interface is built automatically on the fly as Web pages, whereas data are stored in a generic repository. This allows for immediate deployment and population of the database as well as instant online availability of any modification.

Conclusions: OntoCRF is a complete framework to build data repositories with a solid relational storage. Driven by ontologies, OntoCRF is more flexible and efficient to deal with complexity and change than traditional systems and does not require very skilled technical people facilitating the engineering of clinical software systems.

(JMIR Med Inform 2014;2(2):e14) doi:10.2196/medinform.3023

KEYWORDS

biomedical ontologies; data storage and retrieval; knowledge management; data sharing; electronic health records

Introduction

The health sciences, particularly medicine, are based upon information and communication. Clinical practice and research processes consist mostly of collecting data, summarizing this data, and using information derived from the data. This information, properly integrated with clinical knowledge, constitutes the base for decision support and generation of new

knowledge. Nevertheless, in spite of great advances in the information and communication technologies (ICT) domain during past years, the progress in medical informatics is slower than predicted. Clinical information systems are failing to provide true support for clinicians' needs [1,2]. Although there is a broad commercial offer of clinical information systems to support patient management and the electronic patient record (EPR), they are focused primarily on the economic and administrative processes, and lack the needed functionality to manage clinical data. Existing central data warehouses usually fail to support the creation of structured variables for research use [3], so it is necessary to build dedicated systems [4]. As a result, there is little institutional support within health organizations for the collection of clinical data, especially for research.

The implementation of research data repositories has been reported to increase the capacity of a research team [3]. Some surveys show that individual organizations are progressing to the development, management, and use of clinical repositories as a means to support a broad array of research [5]. Although most researchers already use some software system to manage their data, there continues to be widespread use of basic and general-purpose applications, such as spreadsheets, and additional support has become necessary for managing datasets. Interestingly, the barriers to acquiring currently available tools are most commonly related to financial burdens [6].

This is the situation in the Hospital Clinic of Barcelona, which has a long tradition in biomedical research and stands as a benchmark institution both nationally and internationally [7,8]. A research project cannot be understood now without ICT support to some extent. Nevertheless, the spreadsheet remains the key tool for research data management because financial limitations restrict the acquisition of more complex tools. Continuous change is a characteristic of the biomedical domain, and building applications that can handle it is very expensive.

We have developed Onto Clinical Research Forms (OntoCRF), a framework for the definition, modeling, and implementation of data repositories. Most importantly, OntoCRF is capable of meeting change at a minimal cost because the implementation of a new repository in OntoCRF does not need additional database design or programming. All information required to define a new project is explicitly declared in ontologies, reducing the time and cost of development compared to traditional solutions. The repositories implemented with OntoCRF are accessible via a website for data entry, thus facilitating the collection of distributed data.

Methods

Background

The Hospital Clinic of Barcelona has a growing need for systems for the collection of clinical data. The Medical Informatics unit at Hospital Clinic of Barcelona has experience designing and implementing databases for research [9-11]. Some general requirements for data management reported in the literature [3,5,12] are as follows:

1. The ability to efficiently acquire, store, and manage large volumes of structured data, preferably in a centralized repository.
2. To provide a Web interface for researchers to allow them to have a distributed access to the data in order to introduce new data or to retrieve existing data. Data are usually gathered by various researchers, often in different locations.
3. Data security, including access control, to assure the persistence of the data.
4. To facilitate the access to the data, including researcher "self-serve" access.
5. To be able to easily accommodate changes in the structure of the data, minimizing service disruption when such a model change occurs.

The Hospital Clinic of Barcelona has used an EPR system since 1995. Three different commercial systems have been used during this time, the last one including a data warehouse, but they were primarily focused on economic and administrative processes. Although these systems allowed gathering of some limited clinical data, none of them were intended to register additional data.

Because of financial limitations, there has been widespread use by researchers of basic and general-purpose application software, such as spreadsheets. The same situation is reported by other authors [3,6]. The use of general-purpose application software has serious drawbacks: an unfriendly user interface, few guarantees for maintaining the consistency of data, difficulties in sharing and consolidation of data, and limited ability to exploit data. Desktop application software programs are definitively not designed to meet the above mentioned criteria.

When there is an adequate budget available, it is possible to build a more sophisticated system. Usually, these systems are built using a multitier architecture composed of a centralized database, an application server, and a Web server providing the user interface. However, this architecture presents some disadvantages. First of all, the development of such applications is a laborious task, as is their extension to accommodate changes. Consequently, this approach is not suitable for domains where data and model evolution is the norm [12]. Secondly, this classical approach requires a very specialized panel of computer technicians and this often leads to communication problems between the biomedical researchers and the development team. Thirdly, the development cost and the cost of information technology (IT) personnel require a high investment [6] sometimes for a short project time (research projects typically last 2-3 years). Finally, using this kind of approach within a large organization produces applications very different among them, and the distribution of data across multiple sources, which complicates the ability of researchers to use the data for answering their research questions [4].

These considerations—the lack of available tools in our organization and the disadvantages of traditional database systems—prompted us to seek an alternative and to build a platform to deploy research projects and clinical registries.

The advances in knowledge management tools and methodologies in previous years provided the opportunity for

a new approach. Ontologies as explicit conceptualizations of a domain [13] seem well adapted to the task of representing medical data. Ontologies resemble databases from an operational perspective because they can be populated with instance data and deployed as parts of information systems for answering queries [14]. Languages to represent ontologies, such as Ontology Web Language (OWL), are designed to be extensible and able to accommodate model changes. The flexibility of ontologies is a major advantage of the technology [14]. These characteristics make ontologies suitable to build a conceptual platform on which specific applications can be deployed [12].

In addition, the use of ontologies is more and more common in the health care field [15-21], which provides an environment to seamlessly integrate the new information models with existing ontologies.

Use Case Presentation

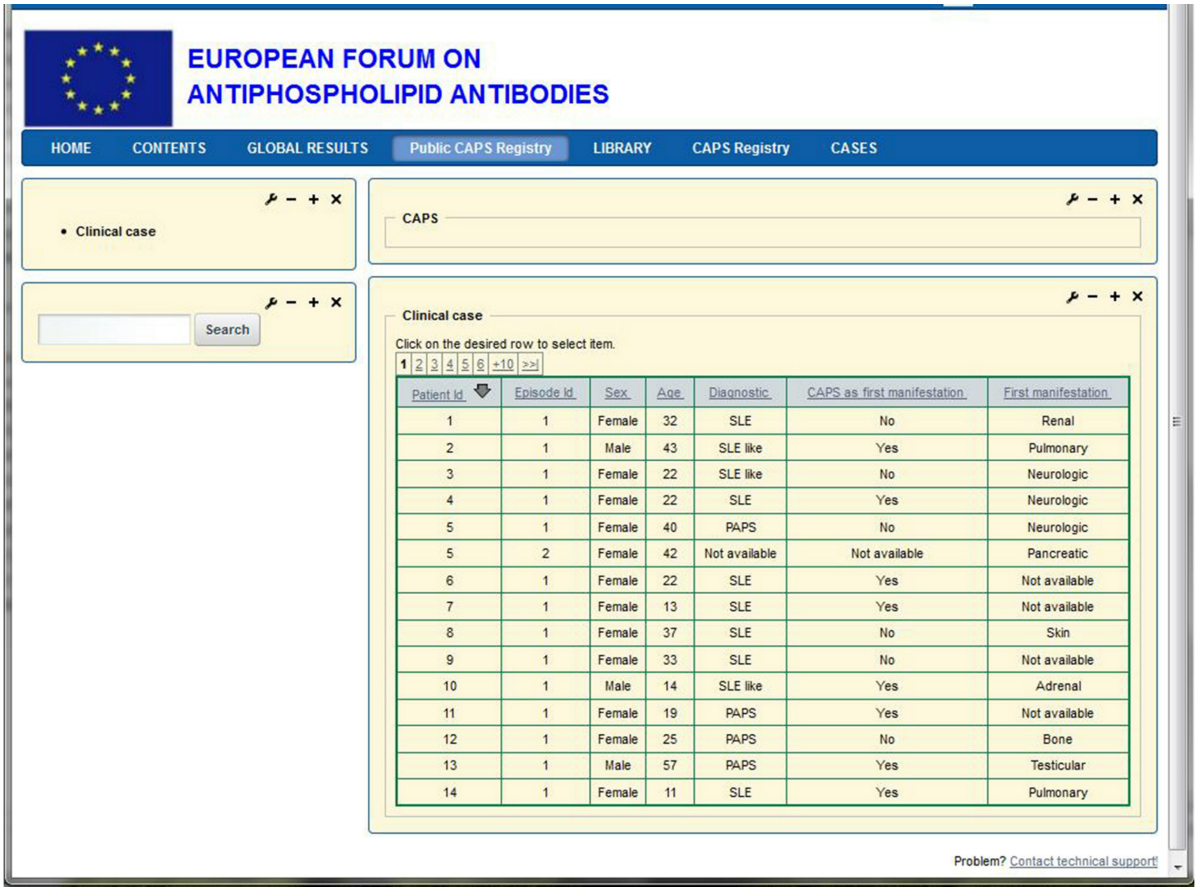
In the following, we will use examples from current projects to illustrate how the system works. One registry is the European Forum on Antiphospholipid Antibodies, a registry of patients with catastrophic antiphospholipid syndrome (CAPS). This project aims to establish an international dataset of all diagnosed patients with CAPS. For each clinical case, the following data are registered: demographic data, previous clinical manifestations, precipitating factors, clinical findings organized by organs, laboratory results, and treatment followed.

Outcome

The data have to be stored in a centralized database to allow periodic statistical analyses on them. In order to allow a decentralized introduction of data, a Web-based application program is needed. Screenshots of the data entry screen are shown in Figures 1 and 2. Figure 1 shows the list of clinical cases from the CAPS registry and Figure 2 shows a concrete case with some laboratory results.

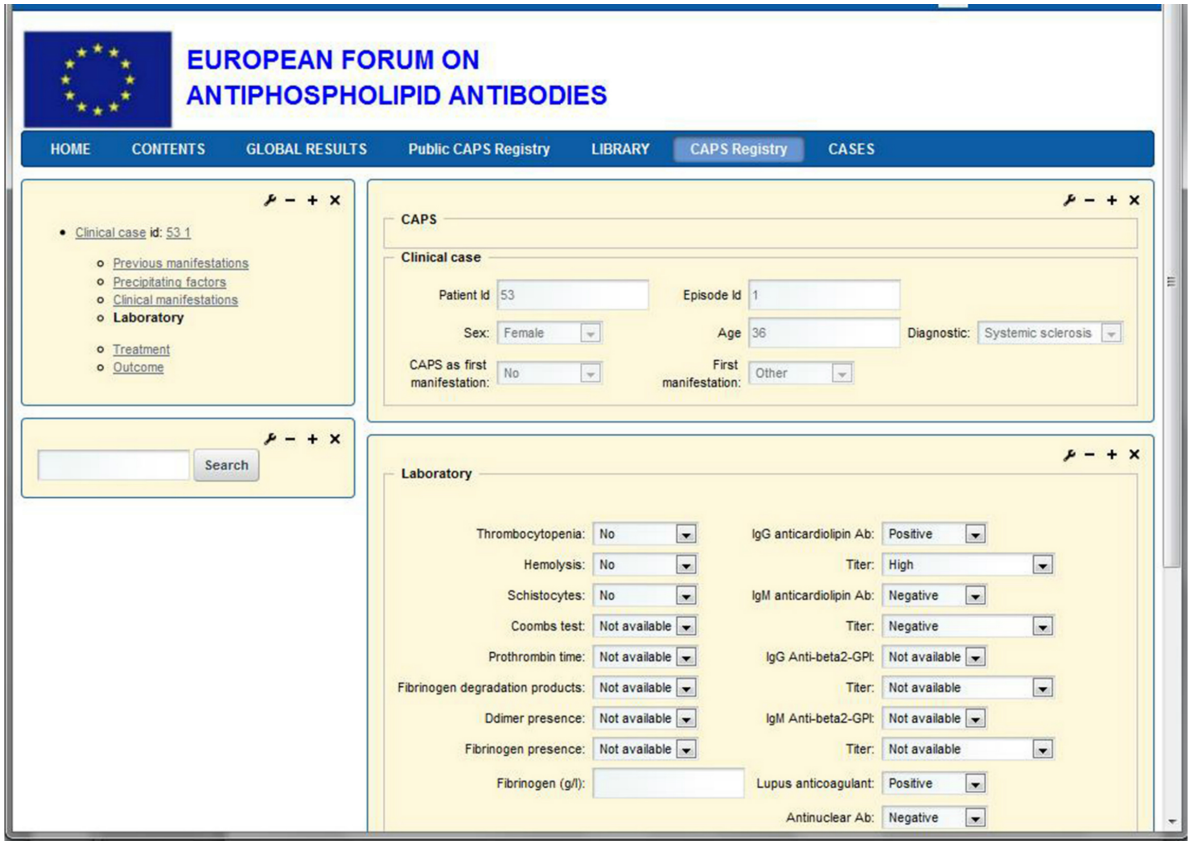
The panel on the top left allows for navigation through the different parts of the registry. The windows on the right, which constitute the formularies to fill in, are composed of single cells, combo boxes, check boxes, radio buttons, etc, to introduce and visualize the data.

Figure 1. List of clinical cases within CAPS registry.



Patient Id	Episode Id	Sex	Age	Diagnostic	CAPS as first manifestation	First manifestation
1	1	Female	32	SLE	No	Renal
2	1	Male	43	SLE like	Yes	Pulmonary
3	1	Female	22	SLE like	No	Neurologic
4	1	Female	22	SLE	Yes	Neurologic
5	1	Female	40	PAPS	No	Neurologic
5	2	Female	42	Not available	Not available	Pancreatic
6	1	Female	22	SLE	Yes	Not available
7	1	Female	13	SLE	Yes	Not available
8	1	Female	37	SLE	No	Skin
9	1	Female	33	SLE	No	Not available
10	1	Male	14	SLE like	Yes	Adrenal
11	1	Female	19	PAPS	Yes	Not available
12	1	Female	25	PAPS	No	Bone
13	1	Male	57	PAPS	Yes	Testicular
14	1	Female	11	SLE	Yes	Pulmonary

Figure 2. A CAPS registry clinical case with laboratory results.



Proposed Solution

OntoCRF is a framework to build clinical data repositories initially designed for research. The general idea of OntoCRF is to combine the best of two technologies: the expressivity and flexibility of ontologies with the proven robustness and efficiency of relational databases. Previous work by our team has already demonstrated the feasibility of using a relational persistence layer to store ontologies [22,23].

As a general requirement, all information needed for the system to work should be modeled in ontologies. Furthermore, no additional programming should be necessary to implement a new project. By doing so, each different project has a different ontology that models both the data and the user interface. The ontology indicates which data are needed (eg, age, sex) and how to represent them on the screen (ie, a single cell in the first row, a radio button in the second row). The program code should be the same for different projects, but is capable of "interpreting" the corresponding ontology to implement different projects.

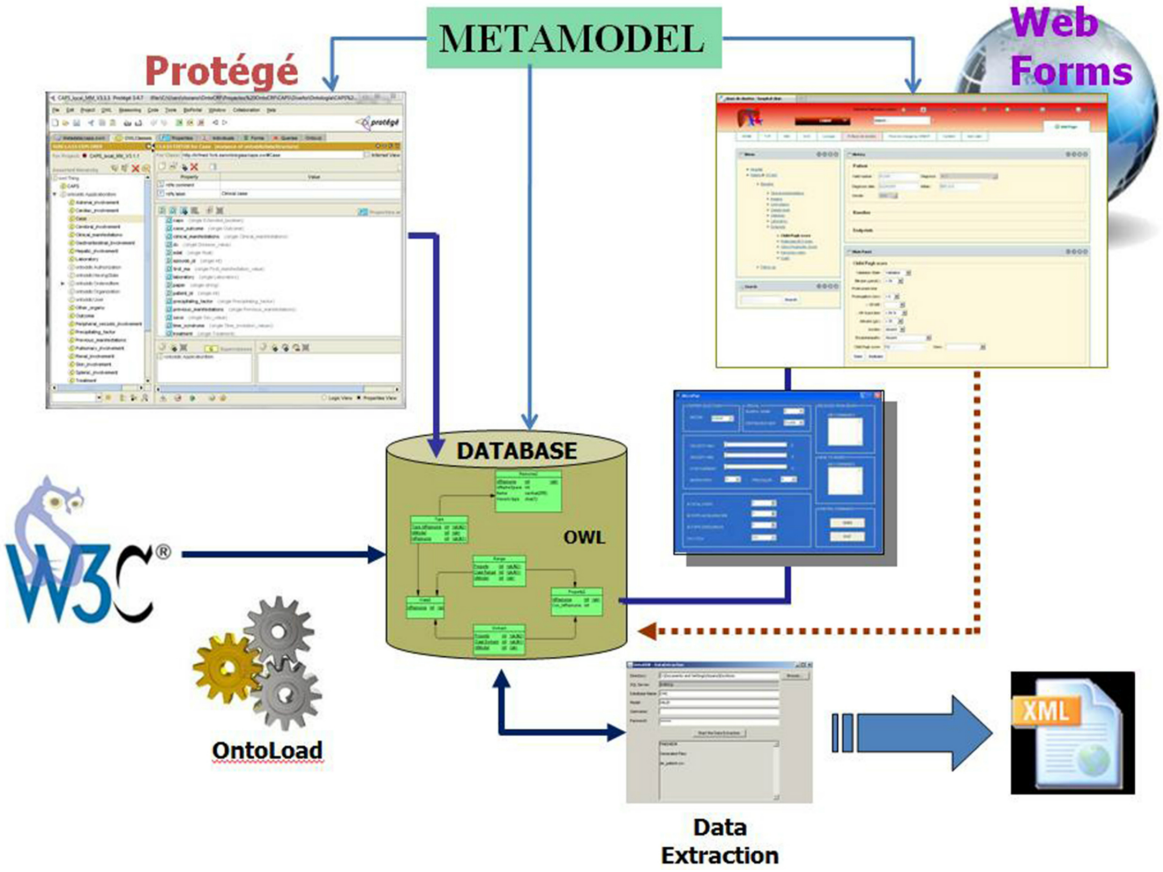
Although prior work was done with Resource Description Framework (RDF), we choose OWL [24] as the modeling language. The justification of using OWL is twofold:

1. Able to reuse existing ontologies. For example, the ontologies stored in BioPortal [25], many in OWL format, are accessible from Protégé.
2. Able to make automatic reasoning in the future. Although not explored yet, we have plans to use reasoners such as Pellet [26] for consistency checking, automatic classification, etc.

OWL is a standard with wide support in the Semantic Web community. Thus, tools developed by the Semantic Web community can be directly applied to the data, such as Protégé [27], as an ontology-editing tool. The election of Protégé is motivated by our previous work on relational support for ontologies [22]. The persistence layer for both models and instantiated data are provided by a relational database.

OntoCRF is composed of the following modules: (1) a relational database for storing the ontologies and instantiated data, (2) an ontology editor based on Protégé, (3) a graphical user interface (GUI) based on Liferay [28], (4) a metamodel describing the primitives of the system, (5) an application for data extraction in the back end, and (6) an application for ontology upload in the back end. The general architecture is shown in Figure 3.

Figure 3. General architecture of OntoCRF.



Storage of Ontologies and Instantiated Data

OWL database (OWL-DB) is a relational database used for storing ontologies and instantiated data, following an approach similar to the Entity-Attribute-Value (EAV) schema. EAV schemas allow for changing the data structure and have proven their utility for clinical applications [29,30] The database was designed according to the OWL specification [24]. Based on Theoharis [31], storage schemes can be classified as schema-oblivious (1 table is used for storing the statements), schema-aware (1 table per class or property is used), and hybrid (1 table per metaclass and property instances with different range values is used).

In OntoCRF, the chosen storage architecture is basically a hybrid model, which is the model that achieves the best performance according to Theoharis [31]. In OWL-DB there is a table for each OWL metaclass, such as resource, class, property, domain, and range. The values of property instances are stored in a table according to its range (eg, resource, string, integer). An identity-based approach is used to identify resources because the use of shorter identifiers versus long internationalized resource identifiers (IRIs) results in space and performance benefits [32].

An additional single table is used to store all triples defining the ontology. Adding or deleting statements in this table causes triggers to fire and thus update the rest of the tables. The statements table serves as interface with other applications. Any application able to manage OWL statements (eg, ontology edition tools) can be potentially connected with OWL-DB.

Furthermore, this approach has all the advantages of EAV schemes. Instead of specific tables for storing patient data, laboratory data, etc, there are tables representing the elements of OWL specification. Therefore, schema evolution can be easily supported. Whereas the addition/deletion of a new property requires the addition/deletion of a table in schema-aware approaches [31], it only requires the addition/deletion of rows in the hybrid model. As a result, neither the design nor the structure of the database needs to be changed for different applications.

The design of the database is intended for a quick recovery of concepts through hierarchies of classes and subclasses. When only using a statements table, finding the subclasses of a class (through a variable number of levels) is a recursive problem, difficult to solve in the relational environment. To avoid this limitation, subsumption relationships between classes and properties are stored in specific tables, following a nested-set model of trees [33]. In this model, each node of the tree is labeled with two numbers (left and right), as shown in Figure 4.

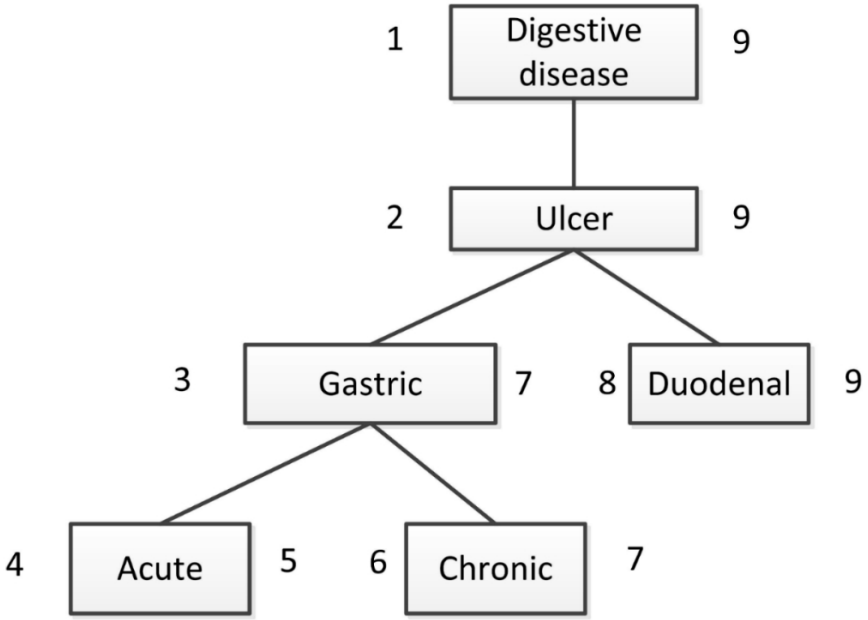
Finding all subclasses of a given class (eg, digestive disease) becomes a very fast process: they are all classes with the right index (or left) comprised between the values of the indexes of the class. Thus, all concepts defined as subclasses of it, such as acute gastric ulcer in the example, will be recovered in a very efficient manner, regardless of what level of depth in the hierarchy they are defined. Nevertheless, this design makes the management of multiple inheritance difficult. Currently, we

duplicate the node with multiple inheritances in the class hierarchy, which represents only a small cost in storage space.

Other applications can interact with OWL-DB using an application programming interface (API) built with stored procedures. A set of functions retrieves the subclasses, properties, and instances of a named class, domain and range of properties, values of instance properties, etc, to extract information from the database.

The system can store all imported ontologies in the same database, maintaining the import relations between different ontologies.

Figure 4. Example of a nested-set model of trees.



Ontology Authoring

The edition of the ontology is based on Protégé [27]. Protégé is a recognized standard for ontology edition, with more than 200,000 registered users around the world, and able to edit OWL ontologies. An interesting characteristic of Protégé is its extensibility capability. It is possible to include new functionalities to the tool by adding new plug-ins.

With OntoCRF, the data to be registered are modeled in an ontology. To simplify and parallel relational databases, tables become classes and columns become properties. Figure 5 shows a snapshot of the CAPS ontology. Some classes representing the main groups of data to be registered (eg, case, Precipitating_Factors, Previous_Manifestations, Adrenal_Involvement, Cardiac_Involvement, laboratory, treatment) can be identified.

OWL and Protégé support additional functionality because the subclasses, metaclasses, etc, together with the metamodel allow Protégé to be used as a twofold design tool: (1) a kind of database design tool to define the data, its structure, and properties and (2) a graphic interface design tool to define how the data will be presented to the user.

A plug-in developed by us, OWL-DB plug-in (Figure 6), connects Protégé with the OWL-DB module at the storage level. The OWL-DB plug-in uses Jena [34] to manage OWL statements and to communicate with OWL-DB.

The plug-in is a backend plug-in. This plug-in consists of a single class, which is subclass of the KnowledgeBaseFactory class provided by Protégé. It communicates by updating the statements table, which triggers the update of the rest of the tables in the database.

By using the OWL-DB plug-in, it is possible to load an ontology that was previously stored in the database to be edited in Protégé. The connection parameters provided are database management system (DBMS), server Internet Protocol (IP) address, database name, username, and ontology namespace. After changes are made in the ontology with Protégé, the user can choose either saving in the database only the last changes made or replacing the ontology entirely. If the ontology is importing other ontologies, an option is available to save all imported ontologies in the database at the same time.

By using the OWL-DB plug-in, an already existing OWL file in Extensible Markup Language (XML) format can be uploaded to the database. This is done using the Protégé menu option “Convert Project to Format...” where an option is available to choose the OWL-DB format. When storing ontologies in OWL-DB from Protégé, a local copy in an OWL file in XML format is automatically generated.

Figure 5. Ontology edition with Protégé.

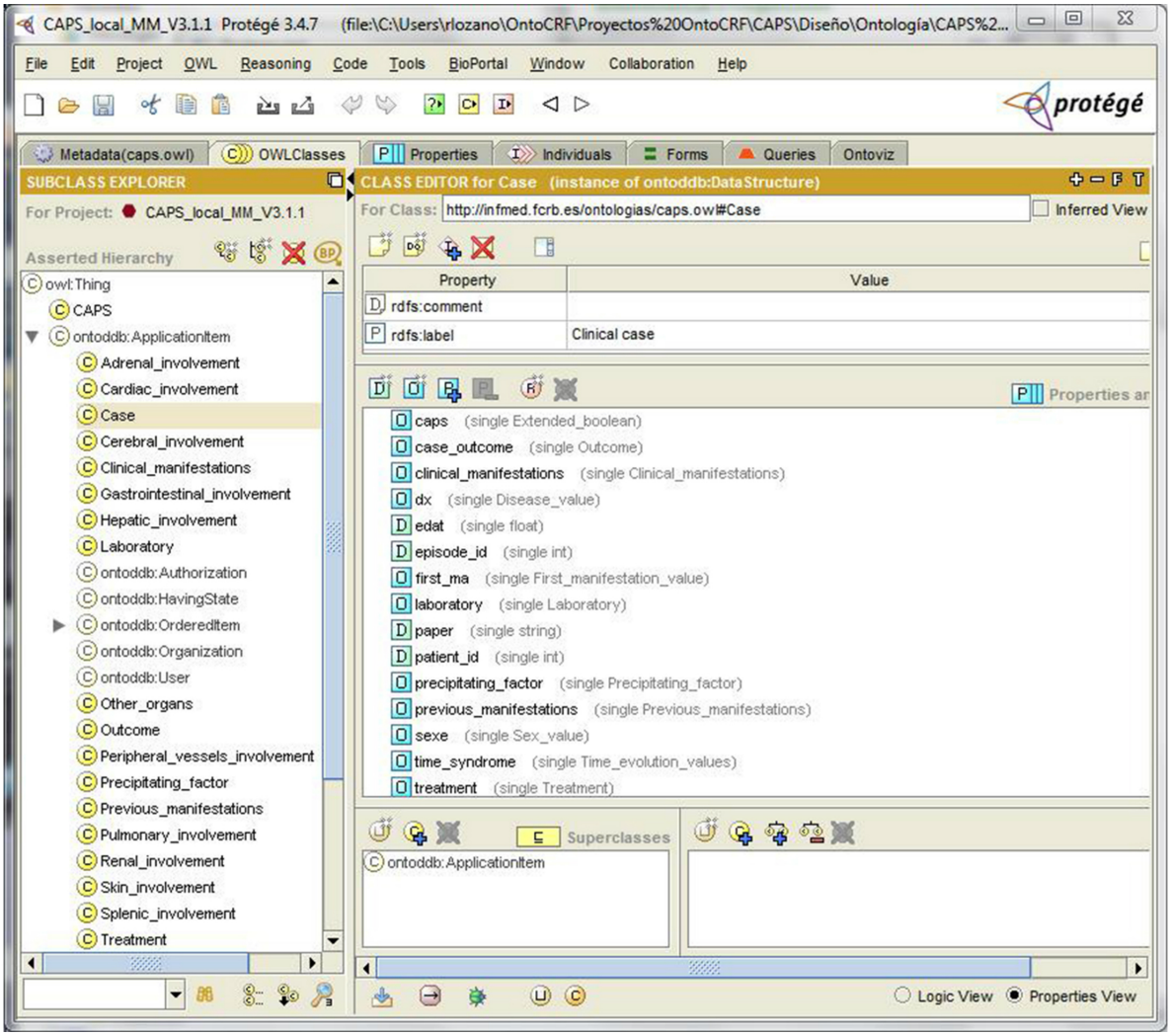
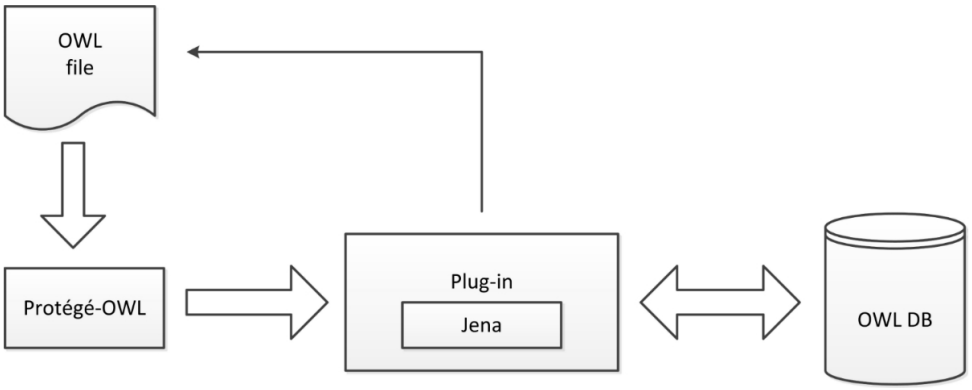


Figure 6. OWL-DB plug-in.



The Metamodel

Ontology-driven database metamodel (OntoDDB-MM), the OntoCRF metamodel, is an ontology composed of a set of metaclasses, classes, and properties that define the available elements that can be used to build an application. These elements are recognized and used by the portlets to create the GUI. Figure 7 shows the main hierarchies.

In this metamodel, an application is represented by an instance of the metaclass application. In the CAPS registry example, the application is represented by the class "CAPS." The different forms are represented by instances of metaclass DataStructure (eg, case, Previous_Manifestations, Clinical_Manifestations). At the same time, these classes are subclasses of the ApplicationItem class.

A DataStructure can have several properties; some are DatatypeProperties and others are ObjectProperties. In our example, the properties Previous_Manifestations, Precipitating_Factors, Clinical_Manifestations, etc, are instances of both ObjectProperties and MenuItem. On the one hand, they

are properties linking Case with other data structures and, on the other hand, they are menu elements. Figure 8 shows an example.

Each form field is an instance of one of the subclasses of FormElement, which determines its behavior: Checkbox, Combobox, Graphic, HyperlinkProperty, ImageProperty, LiteralProperty (to represent literals, do not expect a value), MultilineStringProperty, RadioButton, SingleCell, Password, and SubForm (not implemented yet).

To manage the form fields, the FormElement metaproperty introduces the following facets: webColumn (the relative column in the form where the field will be shown), webRow (the relative row in the form where the field will be shown), webDescriptionProperty (a flag to mark fields that are part of the description of the corresponding object and are shown in the headers, list, etc), webMandatoryProperty (a flag for fields do not allowed to have a null value), webIdProperty (a flag to mark fields that constitutes the Id of the corresponding data structure meaning that is mandatory to fill in the field and that the value must be unique), webEditionDisabled (a flag to avoid a field be edited), and webDirectlyDependent (a flag to identify depending objects). The objects that are values of webDirectlyDependent properties cannot exist without the object that has this property.

The metamodel indicates that there are constraints on the values to be used within each field. This is done by creating a subclass of the class AllowedValues for each field to be constrained. This class is a subclass of OrderedItem and the instances can have a relative order between them. If the subclass CodedValues is used instead of the class AllowedValues, each of the different options can have an attached code. This mechanism is similar to the method used by Rector et al [35] to constrain the codes to placeholders.

As an additional feature, the system can notify to specific users via email about the creation of new instances. This is useful to notify about adverse events, for example. To do that, the class whose instances have to be notified has to be a subclass of the Reportable metaclass.

Other classes, such as Role, UnderAuthorization, Organization, and Authorization, manage access permissions to the different resources, but they are only partially implemented at the present moment.

Figure 7. The ontology-driven database metamodel (OntoDDB-MM).

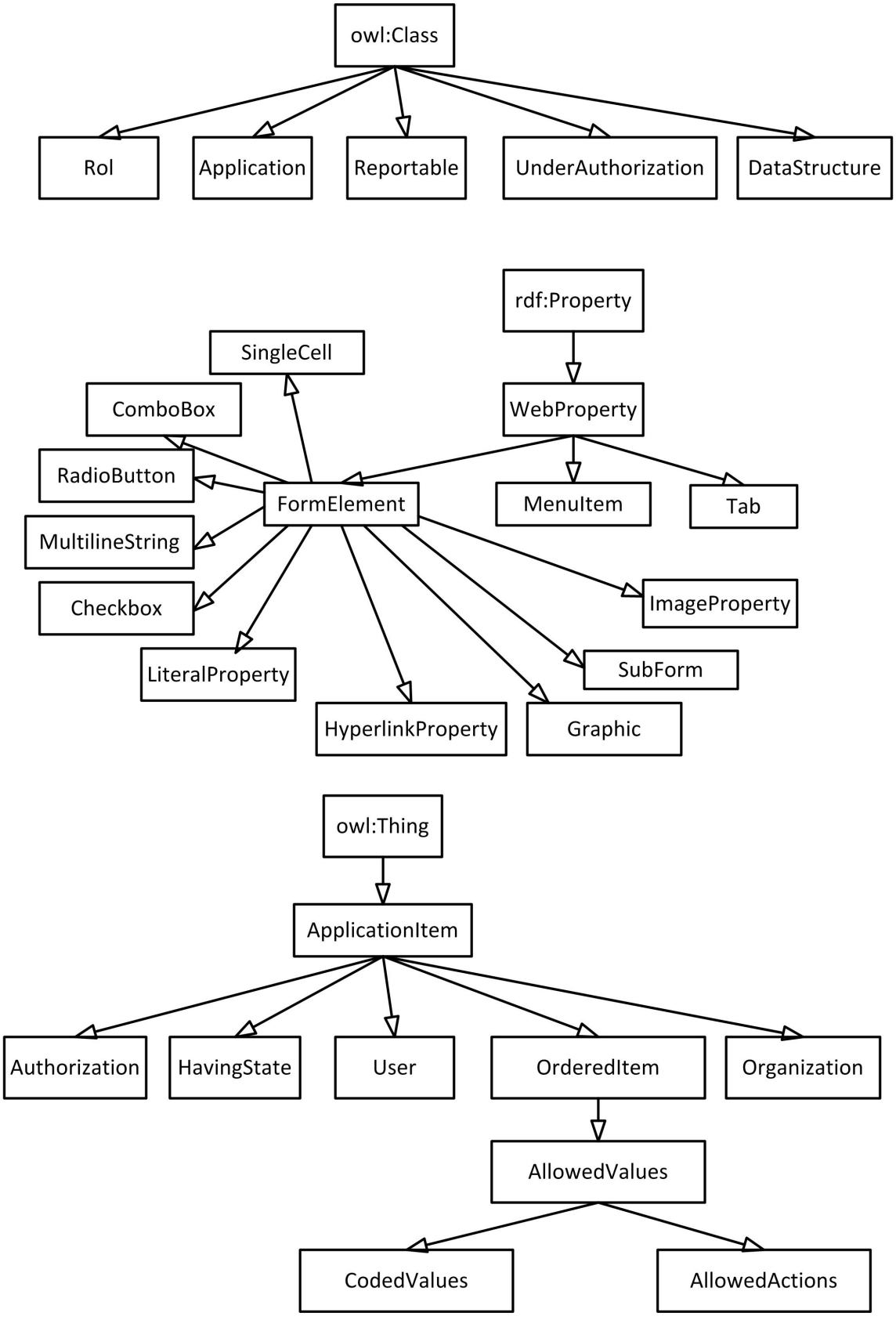
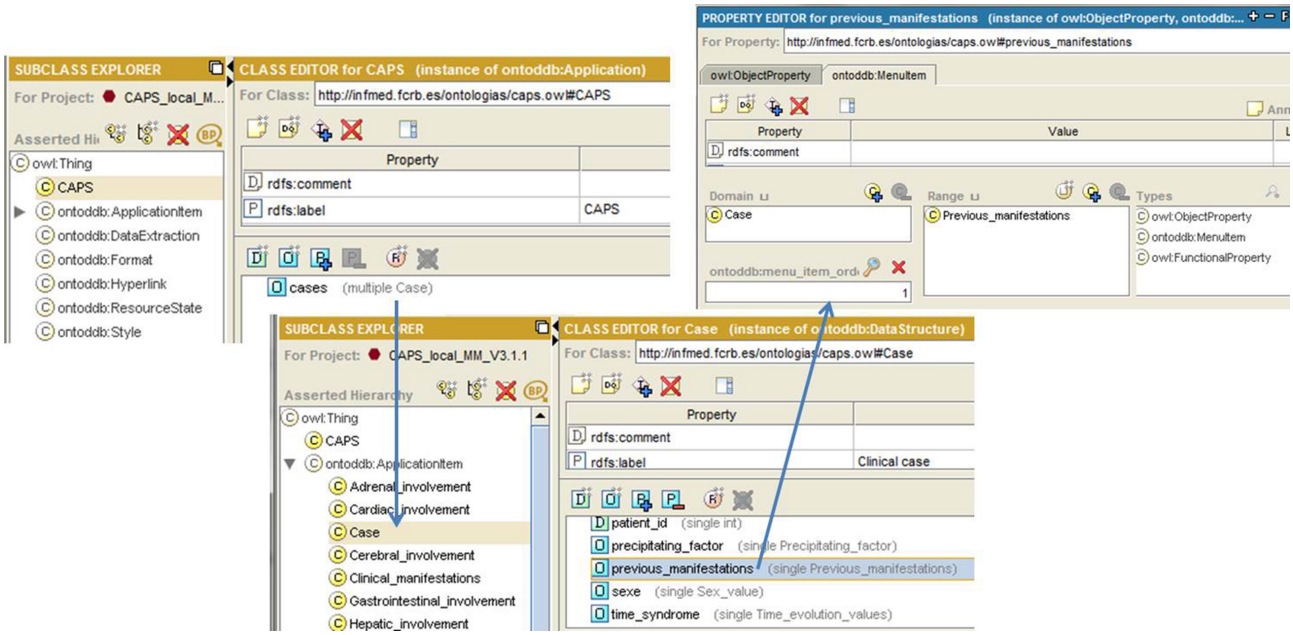


Figure 8. Example of menu elements.



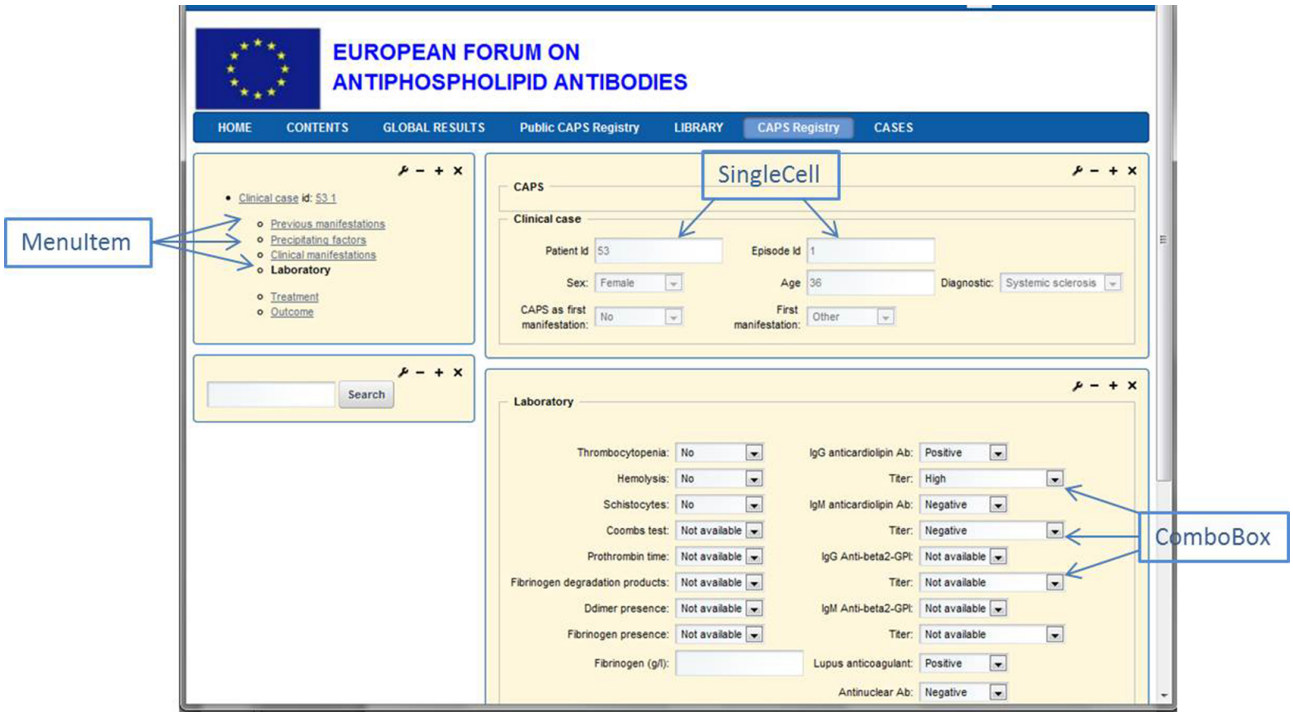
The Graphical User Interface

Using Protégé and OWL-DB is enough to instantiate the ontology in a centralized repository. However, this would not be a suitable interface to an end user.

The user interface is built with portlets based on Spring model-view-controller (MVC) and deployed in Liferay. The business and controller levels are supported by Spring and the view level by JavaServer Pages (JSP) with JSP Standard Tag Library (JSTL). The screen presentation and direct interaction is made with HTML, Javascript, and JQuery. With this approach, the end user only needs a Web browser to interact with the system.

The GUI is created dynamically. The navigation menu, components generation, and all objects in general are created dynamically following the specification of the ontology. The portlets access directly to the OWL-DB stored procedures. Then the information about the application, expressed in the ontology, is used to build the Web pages on the fly, as shown in Figure 9.

Figure 9. Example of form elements.



Data Extraction

The data extraction module allows periodic extractions of stored data for analyzation. This is done by invoking a Java application that will ask the user to provide the connection parameters. The output of this application is a set of XML files containing the data. These files can be imported to a conventional relational database or a statistical package to be analyzed. The data to be extracted is defined in the ontology as instances of the class DataExtraction. This class allows the user to specify which class of the application should be extracted and whether the value of their object properties must be traversed recursively or not.

Another available functionality can transform the entire ontology in a relational database. In this case, the output is a SQL script. This functionality can be used on a daily basis to maintain a relational version of the data.

OWL-DB OntoLoad

OWL-DB OntoLoad is an application to directly upload an OWL ontology in XML format to the server by feeding the statements table directly instead of uploading it through the editor tool.

Evaluation

To evaluate the usability of OntoCRF, the System Usability Scale (SUS) score was selected [36]. Developed in 1986 by Digital Equipment Corporation, it is a simple method to gauge first impressions of the appropriateness of software developments of end users. It consists of a questionnaire with 10 items:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

The answer to each item is a value from 1 and 5 (1=strongly disagree; 5=strongly agree). The results are computed following an algorithm that gives a unique result (SUS score) from 0 to 100.

The data from the questionnaires were entered into a database and analyzed using SPSS 21 statistical package (IBM Corp, Armonk, NY, USA).

Results

OntoCRF has been used in more than 10 different projects. In general, these projects fall into one of the following categories:

1. Research projects with limited duration: a set of data, previously agreed, is collected and analyzed at the end of the project.
2. Clinical registries without a predetermined end date to modify the data collected during the project.
3. Implementation of clinical questionnaires.
4. Nonclinical applications.

The number of cases by project varies between a few hundred to 2000, with approximately 60 to 600 variables per case.

Table 1 shows a summary of the characteristics of the main projects running currently. The upload and download was made between Protégé 3.5 and OWL-DB, and refers to the entire ontology. To measure upload and download times without being influenced by traffic on the Internet, a local server was used with the following characteristics: SUSE Linux Enterprise Server 11 (x86_64) operating system, GNU/Linux 2.6.32.43-0.4-default x86_64 1 x Intel Xeon CPU E5-4640 0 @ 2.40GHz CPU, and 4Gb RAM.

Table 1. Summary of characteristics of the projects implemented.

Project	Number of statements	Disc space (Mb)	Number of classes	Number of properties	Number of instances	Upload time (sec)	Download time (sec)
1	102,371	48	147	365	26,414	65	12
2	103,652	72	91	288	8794	58	24
3	191,487	90	81	171	26,408	112	24
4	200,509	98	15,982	90	15,595	311	80
5	264,636	126	258	632	32,317	200	27
6	131,926	74	145	623	7553	75	17
7	79,350	47	125	535	5378	44	9

In all projects, OntoCRF has been able to meet their specific requirements and to cope with the requirements of modifications during the lifecycle of the projects. The modular architecture of the metamodel has proven its feasibility to accommodate new extensions of the system. Also, the separation of data layer and presentation layer allows the progressive addition of new functionalities as needed.

The flexibility provided by the system facilitates to provide prototypes from the initial moment, which is a very valuable resource for developers to work close to the physicians. From the beginning of the project, key users have material to work

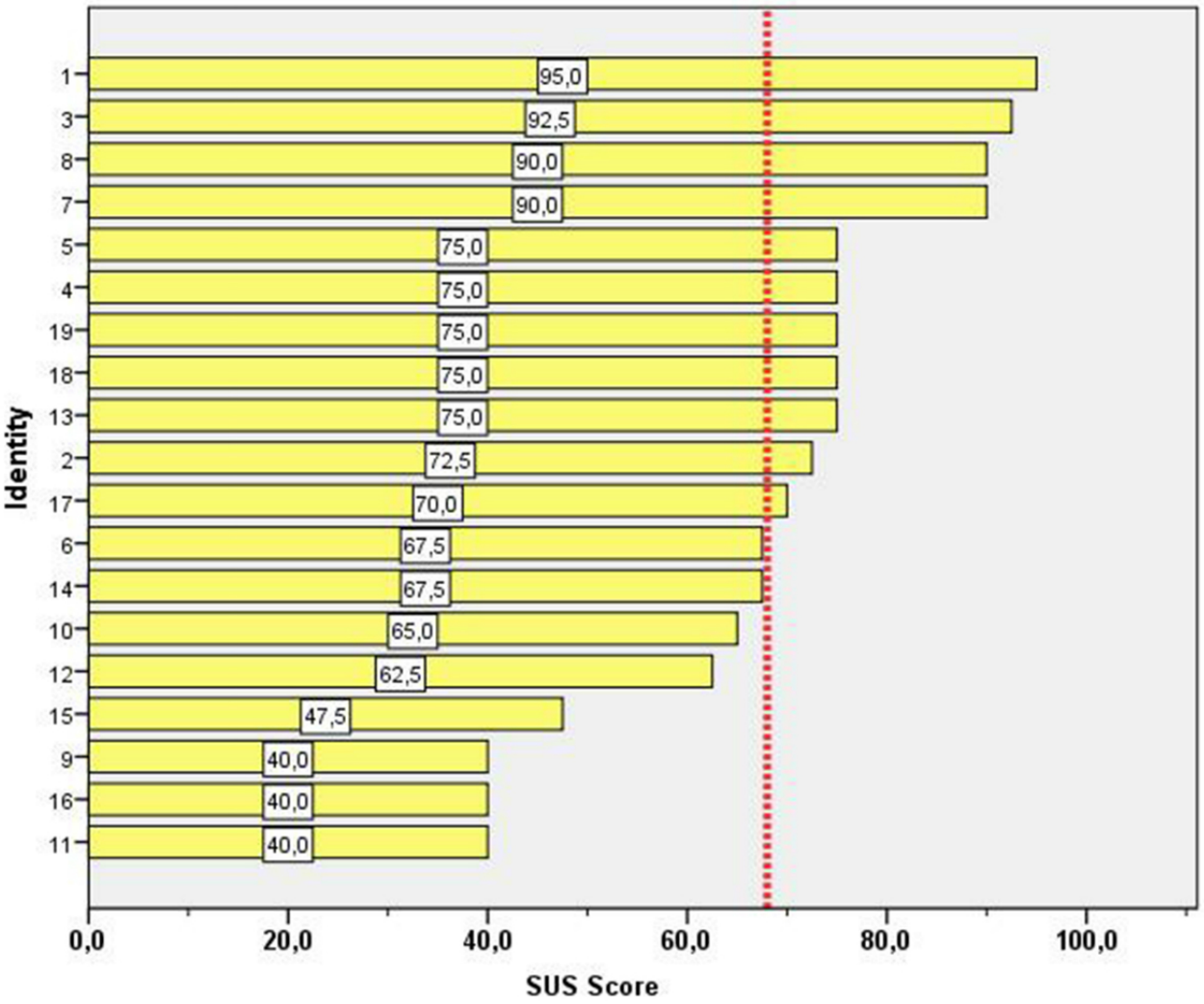
with, and it is possible to make online modifications and check results immediately.

A survey was distributed to a sample of 35 OntoCRF active users who used the system on a daily basis. Of these 35, 19 (54%) answered the questionnaire. Data were introduced into a database and the SUS score was computed. The results are displayed in Figure 10.

Of the 19 respondents, 11 (58%) computed a global SUS score greater than 68 which is recognized as "above average" [37]. According to Bangor et al [38], it is possible to grade over a curve based on the distribution of all scores in relationship with their quartile position. In all, 4 users (21%) gave the solution an A grade (excellent), 5 (26%) gave a C grade (good), 6 (32%) gave a D grade (pass), and 4 (21%) rated the solution with an F grade (fail) [38]

Because of the success achieved with OntoCRF in the first projects, which were primarily research projects, OntoCRF is currently being marketed and used in new types of projects.

Figure 10. Results of the computed SUS score by respondent. Results are displayed in ascending order. The dotted line marks a score of 68 (above average).



Discussion

Overview

The focus of OntoCRF is to assist with data collection during research studies, automating the process as much as possible and minimizing the technical knowledge required from the final users for the creation and management of new studies. In particular, we provide an automatic system for dynamic creation of Webs driven by ontologies and with additional tools for the extraction and analysis of the data.

Our system, unlike other solutions, does not work with triples or RDF graphs; it works with ontologies, particularly those represented in OWL. Ontologies are stored in a relational database directly in OWL, following a hybrid approach. This eases the querying process because there is no OWL-SQL mapping needed. For instance, to retrieve the classes, the system just accesses the "class" table. Further logic is not necessary and it can all be done through simple SQL queries. Because we have to deal with very large ontologies, performance was a critical feature from the beginning; this OWL-driven approach achieved our efficiency requirements, whereas other systems failed.

Regarding the performance of the system, its behavior is quite linear. As Table 1 shows, any of the variables considered has a preponderant influence. In general, the upload and download times are proportional to the number of statements. The greater complexity of some ontologies, expressed by a higher proportion of classes and properties in relation to the number of instances, involves a slight penalty. Project 4 (with 2 orders of magnitude more in number of classes) showed a worse performance, but less than 4 times worse than other projects with a similar number of statements. This is due to the cost of maintaining the class hierarchy tree in the database, primarily when uploading the ontology. In previous versions of the system, each time a class was inserted in the database, all the indexes of the class hierarchy were recalculated. Project 4 showed the lack of scalability and efficiency of this approach; the system was not able to recalculate the indexes and remained working without end. In the current version, the entire class hierarchy is calculated only once after all classes have been inserted into the appropriate table. This approach represents only a gain of 2-3 seconds for the rest of the projects (not shown in the table), but a radical change for projects with a large number of classes. With this approach, the cost of maintaining the class hierarchy is assumable; in return, the retrieval of instances at whatever level is trivial.

The previous discussion is about uploading and downloading the entire ontology, a task that is performed during the development phase of a project. The user interaction with the system, adding and retrieving data, is no different from other systems. The user interaction involves only a small set of data, not the entire ontology.

OntoCRF demonstrates that an ontology-based approach is more flexible and efficient to deal with complexity and change than a traditional system, facilitating the engineering of clinical software systems. First of all, the application development phase is reduced to only analysis and design. The availability of prototypes from the very beginning, and the facility to apply changes, make OntoCRF an extremely useful tool to check the requirements and the solutions proposed. These facts imply a very important drop in costs and time with their consequent savings.

Secondly, differences between applications are reduced to their conceptual model. Therefore, the same infrastructure can be used for different projects, taking advantage of scale economy. All projects implemented until now share the same hard and soft infrastructure. The only difference between them is the content.

At the conceptual level, some elements or models can be reused in different projects, so homogeneous criteria and conceptual models could be established inside an organization. Concepts such as patient, clinical manifestations, and laboratory results are common in different projects, so these definitions can be easily shared and extended as needed.

The use of ontologies provides the ability to manage data structures declaratively, thus focusing the design on the conceptual aspects and not on the technical issues. Making an ontological analysis of an application allows for focus on a higher abstraction level and to concentrate on the domain aspects, thus helping researchers to clarify the implicit knowledge to manage. Moreover, the communication between designers and users is established at a conceptual level. Technical discussions that often contaminate the conceptual analysis in other approaches can be avoided. Moreover, ontologies assure that data and knowledge used in the project remain well documented.

Because the solution allows for modification of the underlying schema of the data, some measures are needed to guarantee the consistency of the instances. Problems could arise if trying to modify or delete classes or properties. The first security level is provided by Protégé, which does not allow performing some actions that could leave the ontology in an inconsistent state. This is the case when trying to delete a class that has instances. The rest of cases should be solved by the specification of editorial policies. When a project is running, deleting a class or property could be replaced by setting a deprecated flag on the resource. Nevertheless, in the database data are never physically deleted, only a delete flag is used to prevent the loss of data by mistake.

We consider the use of OWL and Protégé a good choice. The expressivity power of the language was adequate to cover the requirements of the projects in which OntoCRF was used. Moreover, it eases the interchange and reuse of models. The use of OWL allows adding reasoning capabilities in the future, a very promising line to explore.

From the usability study, it can be concluded that OntoCRF is well accepted by nearly 60% of its users, who considered the solution globally above average. But in a more detailed look at the data, high fragmentation is shown resulting in 4 groups with a very different perception of usability, from the best grade of “excellent” to the worst as “fail.” One explanation for such discrepancy could be a misunderstanding of the product under evaluation. OntoCRF has 2 components: a portal (developed using Liferay) customizable by the administrator of each community, and a database access for collecting the data. Moreover, OntoCRF is conceived as a full service in the cloud. Therefore, many different factors and user experiences can be interposed in the routine operation. The SUS score was developed in 1986, when many software solutions were developed for mainframe use or in a client-server environment. At present, widespread Internet usage interposes many more layers between the user interface and the physical data repositories. In this scenario, we need to better inform the users about what is being measured with the SUS score tool and perhaps develop new tools better suited for such new systems architecture. Nevertheless, further usability studies are required to improve OntoCRF, including specific questions with better information about the reasons of a low grading by some users.

Although the system is primarily used in health-related projects, the model is totally independent of the domain, so it would be suitable to gather data in any context. In fact, some projects implemented with OntoCRF are not about clinical information, but about management-related data. In general, if it is possible to model the data with OWL, it is possible to use OntoCRF.

Limitations

We are aware of the system limitations. The metamodel of OntoCRF is not capable of process representation; hence, it is not able to manage explicit knowledge related to processes at the moment. The data extraction capacity is also limited. Currently, the final user cannot perform direct consultations over the server. Instead, data need to be previously extracted. This limitation is currently being addressed and some tools are being tested with the aim to be integrated with OntoCRF.

Comparison With Prior Work

OntoCRF proposes the use of ontologies to ease and speed up the development of data repositories. The ontology-driven development of complex and intelligent systems has been largely applied in the past, especially when the ontologies or the methods are likely to be reused for new or derivative applications [16,39,40]. In general, the goal is to transform the system development cycle, so instead of programming each new application from scratch, we can select, modify, and assemble existing components [41]. Ontologies are used to build knowledge bases containing detailed descriptions of particular application areas. OntoCRF goes a step further because there is no need for programming, just the design of the application ontology. OntoCRF ontologies contain not only knowledge about the domain, but also the detailed description of the application.

The discussed approach is also related to the model-driven architecture (MDA) launched by the Object Management Group (OMG) [42]. According to their manifesto, MDA is a style of enterprise application development and integration based on using automated tools to build system-independent models and transform them into efficient implementations. As with ontology-oriented approaches, software evolution is handled simply by editing the underlying model. OMG is guided to object-oriented applications, particularly to distributed ones. It represents a more technical approach, centered on the platform independence, whereas OntoCRF pursues the conceptual independence. In our case, the database never changes and neither does the implementation of the application. Our work represents an advance because everything is defined explicitly, but the use is much more restricted.

In regard to research in data repositories, there exist multiple ontology-driven solutions for discovering and searching existing resources [43,44] or to consolidate clinical research data from disparate databases [4], but not much for automatically building new ones.

Compared with Protégé, WebProtégé [45] adds collaboration support and improves knowledge acquisition, but remains primarily an ontology editor. The work of Li et al [12] is close to our work in considering ontologies as the center of the architecture. The proposed system is focused on modeling a domain and supporting data and model changes, through versioning and dynamic composition, while using a simple interface with few options. On the other hand, Butt et al [46] propose the automatic generation of Web forms from ontologies with the objective of facilitating the creation of RDF data. Although the system produces easy-to-use forms, the capabilities of structuring the information are very limited.

As of this writing and to the best of our knowledge, there are no frameworks allowing the creation of data repositories, with the interface functionalities of traditional systems, in such a dynamic way such as OntoCRF, where even the user interface is built through the edition of ontologies.

There exist several works regarding how to store RDF graphs and ontologies, such as triple stores or relational databases that may be accessed as RDF graphs. However, none of these fulfill the needs of our system. In the case of RDF-based access to relational databases, such as the platform D2RQ [47], the system is read-only and just provides a RDF view of the content, but it does not provide any solution for storing the content, instead relying on an existing database created by the user. Also, the user has to generate the mappings between the platform and the database, specific for each use case. In the case of triple stores, they offer a way to store and retrieve triples, leaving the logic necessary for interpreting the triples and retrieving the right ones to an API or a query engine. This requires analyzing the whole set of triples of a specific graph, which has a high cost and is not scalable to big ontologies.

Our repository is not the only one with these characteristics, although it was at the time of our search for solutions. Systems such as OWLIM [48] and DLDB2 [49] combine DBMSs with additional capabilities for partial OWL reasoning. Furthermore, there are repositories with similar architectures to ours, such as Minerva [50] repository within the Integrated Ontology Development Toolkit by IBM. Because OntoDDB does not have Simple Protocol and RDF Query Language (SPARQL) capability yet, we could not perform any reliable comparison under equivalent conditions to these similar repositories.

Future Work

We are considering different lines for future work. As previously mentioned, we are currently working on the integration of existing data query tools with OntoCRF to provide query functionalities to the final user. We plan to include SPARQL in the following months.

In a different line of work, we plan to use OntoCRF as the framework to build new electronic medical record (EMR) systems semantically interoperable. OWL representation provides an environment to integrate information models and terminology models used in the clinical context [35]. Currently, we have a prototype which implements the standard ISO 13606 in a native way, and there is ongoing work to conform to the standard EN 13940. These solutions were built using OntoCRF.

Finally, promising research work is being done to use existing ontologies and tools more intensively. Currently, ontologies are being used in OntoCRF as a data-modeling tool, so the use of already existing ontologies is a natural step. Moreover, applying automatic reasoning to data gathered in a project and integrated with external ontologies could provide interesting benefits.

Conclusions

OntoCRF is a complete framework to build data repositories because it includes design of the system, storage, and GUI. The

combination of ontologies and relational technology provides a system that is both flexible and solid. The ontology-based approach is more flexible and efficient to deal with complexity and change than traditional systems. On the other hand, storing the data in a relational database provides the known advantages of a solid relational model.

Although the GUI was not among our priorities, most participants of our usability study computed a global SUS score over 68, which is recognized as above average.

OntoCRF does not require very skilled technical people to make a new project, easing the engineering of clinical software systems. Moreover, the reduction of the development phase implies an important drop in costs and time. Furthermore, because the same infrastructure can be used for different projects, there is no need to dedicate specific equipment for each new project.

At the conceptual level, the ontological analysis of applications allows for concentration on the domain aspects, helping researchers to clarify the implicit knowledge to manage and to facilitate the communication between designers and users. Because some concepts are common in different projects, the models can be reused. On the other hand, ontologies assure that data and knowledge used in the project remain well documented. In addition, OWL and Protégé have proven enough expressivity to cover the requirements of the projects in which OntoCRF was used.

Finally, although currently the system is primarily used in health-related projects, the model is independent of the domain and can be useful in any project in which a distributed collection of data is needed.

Acknowledgments

The authors wish to thank Mark Musen for his comments on the drafts of the paper.

Conflicts of Interest

Xavier Pastor and Raimundo Lozano-Rubí are employees of Hospital Clinic of Barcelona and University of Barcelona, which receive royalties from a third party by the commercialization of OntoCRF.

References

1. Ball MJ, Silva JS, Bierstock S, Douglas JV, Norcio AF, Chakraborty J, et al. Failure to provide clinicians useful IT systems: opportunities to leapfrog current technologies. Methods Inf Med 2008;47(1):4-7. [Medline: 18213422]
2. Lehmann CU, Altuwaijri MM, Li YC, Ball MJ, Haux R. Translational research in medical informatics or from theory to practice. A call for an applied informatics journal. Methods Inf Med 2008;47(1):1-3. [Medline: 18213421]
3. Hruby GW, McKiernan J, Bakken S, Weng C. A centralized research data repository enhances retrospective outcomes research capacity: a case report. J Am Med Inform Assoc 2013 May 1;20(3):563-567 [FREE Full text] [doi: 10.1136/amiajnl-2012-001302] [Medline: 23322812]
4. Cimino JJ, Ayres EJ. The clinical research data repository of the US National Institutes of Health. Stud Health Technol Inform 2010;160(Pt 2):1299-1303 [FREE Full text] [Medline: 20841894]
5. MacKenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. J Am Med Inform Assoc 2012 Jun;19(e1):e119-e124 [FREE Full text] [doi: 10.1136/amiajnl-2011-000508] [Medline: 22437072]
6. Anderson NR, Lee ES, Brockenbrough JS, Minie ME, Fuller S, Brinkley J, et al. Issues in biomedical research data management and analysis: needs and barriers. J Am Med Inform Assoc 2007;14(4):478-488 [FREE Full text] [doi: 10.1197/jamia.M2114] [Medline: 17460139]
7. Asenjo MA, Bertrán MJ, Guinovart C, Llach M, Prat A, Trilla A. [Analysis of Spanish hospital's reputation: relationship with their scientific production in different subspecialities]. Med Clin (Barc) 2006 May 27;126(20):768-770. [Medline: 16792980]
8. Ranking de resultados de los INSTITUTOS DE INVESTIGACIÓN SANITARIA. URL: http://www.isciii.es/ISCIII/es/contenidos/fd-el-instituto/fd-comunicacion/fd-noticias/21Dic2012-Institutos-de-Investigacion-Sanitaria.shtml [accessed 2013-10-08] [WebCite Cache ID 6KDY23F8v]
9. Lozano-Rubi R, Prat S, Echeverría T, Pastor X. Trauma registry. 1999 Presented at: Third European Conference on Electronic Health Care Records; May 6-7, 1999; Seville p. 199-208.
10. Lozano-Rubi R. Registros Clínicos electrónicos y su explotación para la investigación. Todo Hospital 1999(162):817-822.
11. Pastor X, Lozano-Rubi R. Representación de la información clínica e investigación sobre salud. In: Informática Biomédica. Madrid: INBIOMED; 2004:115.
12. Li YF, Kennedy G, Ngoran F, Wu P, Hunter J. An ontology-centric architecture for extensible scientific data management systems. Future Generation Computer Systems 2013 Feb;29(2):641-653. [doi: 10.1016/j.future.2011.06.007]
13. Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisition 1993 Jun;5(2):199-220. [doi: 10.1006/knac.1993.1008]
14. Neuhaus F. Towards ontology evaluation across the life cycle. Applied Ontology 2013;8(3):179-194.

15. Smith B, Brochhausen M. Putting biomedical ontologies to work. Methods Inf Med 2010;49(2):135-140 [FREE Full text] [doi: 10.3414/ME9302] [Medline: 20135080]
16. Knublauch H. Ontology-driven software development in the context of the semantic web: an example scenario with Protégé/OWL. In: Proceedings of the 1st International Workshop on the Model-Driven Semantic Web. 2004 Presented at: 1st International Workshop on the Model-Driven Semantic Web; September 21, 2004; Monterey, CA.
17. Senger C, Seidling HM, Quinzler R, Leser U, Haefeli WE. Design and evaluation of an ontology-based drug application database. Methods Inf Med 2011;50(3):273-284. [doi: 10.3414/ME10-01-0013] [Medline: 21057721]
18. Soguero-Ruiz C, Lechuga-Suárez L, Mora-Jiménez I, Ramos-López J, Barquero-Pérez Ó, García-Alberola A, et al. Ontology for heart rate turbulence domain from the conceptual model of SNOMED-CT. IEEE Trans Biomed Eng 2013 Jul;60(7):1825-1833. [doi: 10.1109/TBME.2013.2243147] [Medline: 23372067]
19. Bodenreider O, Burgun A. Biomedical ontologies. In: Chen H, Fuller S, Hersh WR, Friedman C, editors. Medical Informatics: Knowledge Management and Data Mining in Biomedicine. New York: Springer; 2005.
20. Haug PJ, Ferraro JP, Holmen J, Wu X, Mynam K, Ebert M, et al. An ontology-driven, diagnostic modeling system. J Am Med Inform Assoc 2013 Jun;20(e1):e102-e110 [FREE Full text] [doi: 10.1136/amiajnl-2012-001376] [Medline: 23523876]
21. Tao C, Jiang G, Oniki TA, Freimuth RR, Zhu Q, Sharma D, et al. A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data. J Am Med Inform Assoc 2013 May 1;20(3):554-562 [FREE Full text] [doi: 10.1136/amiajnl-2012-001326] [Medline: 23268487]
22. Lozano-Rubi R, Geva F, Saiz S. Relational support for Protégé. 2003 Presented at: Sixth International Protégé Workshop; July 7-9, 2003; Manchester.
23. Lozano-Rubi R, Pastor X, Lozano E. OntoDDB - Ontology Driven Database. In: Proceedings of the First Symposium on Healthcare Systems Interoperability. 2009 Presented at: First Symposium on Healthcare Systems Interoperability; April 2009; Alcalá de Henares (Madrid) p. 31-38.
24. OWL Web Ontology Language Reference. URL: http://www.w3.org/TR/owl-features/ [accessed 2013-10-08] [WebCite Cache ID 6KDYDWVMT]
25. Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story MA, NCBO team. The National Center for Biomedical Ontology. J Am Med Inform Assoc 2012;19(2):190-195 [FREE Full text] [doi: 10.1136/amiajnl-2011-000523] [Medline: 22081220]
26. Clark&Parsia. Pellet: OWL 2 Reasoner for Java URL: http://clarkparsia.com/pellet [accessed 2014-04-15] [WebCite Cache ID 6OqovxHSp]
27. protege. URL: http://protege.stanford.edu/ [accessed 2013-10-08] [WebCite Cache ID 6KDYHWgrU]
28. Liferay. URL: http://www.liferay.com/ [accessed 2013-10-08] [WebCite Cache ID 6KDYN1Ey3]
29. Anhøj J. Generic design of Web-based clinical databases. J Med Internet Res 2003 Nov 4;5(4):e27 [FREE Full text] [doi: 10.2196/jmir.5.4.e27] [Medline: 14713655]
30. Johnson SB, Paul T, Khenina A. Generic database design for patient management information. Proc AMIA Annu Fall Symp 1997:22-26 [FREE Full text] [Medline: 9357581]
31. Theoharis Y, Christophides V, Karvounarakis G. Benchmarking database representations of RDF/S stores. Lecture Notes in Computer Science 2005;3729:685-701. [doi: 10.1007/11574620_49]
32. Das S, Srinivasan J. Database technologies for RDF. Lecture Notes in Computer Science 2009;5689:205-221. [doi: 10.1007/978-3-642-03754-2_5]
33. Celko J. Nested set model of trees in SQL. In: Joe Celko's SQL for Smarties: Advanced SQL Programming. San Francisco: Morgan Kaufmann; 1999.
34. Apache Jena. URL: http://jena.apache.org/ [accessed 2013-10-08] [WebCite Cache ID 6KDYSw2MP]
35. Rector AL, Qamar R, Marley T. Binding ontologies and coding systems to electronic health records and messages. Applied Ontology 2009;4(1):51-69. [doi: 10.3233/AO-2009-0063]
36. Brooke J. A quick and dirty usability scale. In: Usability Evaluation in Industry. London: Taylor & Francis; 1996:189-194.
37. Sauro J. Measuring Usability. 2011 Feb 02. Measuring Usability with the System Usability Scale (SUS) URL: http://www.measuringusability.com/sus.php [accessed 2014-07-15] [WebCite Cache ID 6R5MsDupd]
38. Bangor A, Kortum P, Miller J. Determining what individual SUS Scores mean: adding and Adjective Rating Scale. Journal of Usability Studies 2009;4(3):114-123.
39. Musen MA. Ontology-oriented design and programming. In: Knowledge Engineering and Agent Technology. Amsterdam: IOS Press; 2004.
40. Musen MA, Schreiber AT. Architectures for intelligent systems based on reusable components. Artif Intell Med 1995 Jun;7(3):189-199. [Medline: 7581622]
41. Musen MA. Scalable software architectures for decision support. Methods Inf Med 1999 Dec;38(4-5):229-238. [doi: 10.1267/METH99040229] [Medline: 10805007]
42. Booch G, Brown A, Iyengar S, Rumbaugh J, Selic B. An MDA manifesto. In: The MDA Journal: Model Driven Architecture Straight from the Masters. Tampa, FL: Meghan-Kiffer Press; 2004.
43. Haendel M, Torniai C, Vasilevsky N, Hoffmann S, Bourges-Waldegg D. eagle-i: ontology-driven federated search and data entry tools for discovering biomedical research resources. In: Proceedings of the 4th International Conference on

Biomedical Ontology. 2013 Jul Presented at: 4th International Conference on Biomedical Ontology; July 7-9, 2013; Montreal, Canada.
44. DataCite. URL: http://www.datacite.org/ [accessed 2013-10-08] [WebCite Cache ID 6KDYXSpD7]
45. Tudorache T, Nyulas C, Noy NF. A distributed ontology editor and knowledge acquisition tool for the web. Semantic Web Journal 2011 Jan 01;4(1):89-99. [doi: 10.3233/SW-2012-0057]
46. Butt AS, Haller A, Liu S, Xie L. ActiveRaUL: Automatically Generated Web Interfaces for Creating RDF Data. Semantic Web 2013:2013.
47. D2RQ: Accessing Relational Databases as Virtual RDF Graphs. URL: http://d2rq.org/ [accessed 2014-04-15] [WebCite Cache ID 6Oqojlsnb]
48. Bishop B, Kiryakov A, Ognyanoff D, Peikov I, Tashev Z, Velkov R. OWLIM: A family of scalable semantic repositories. Journal Semantic Web 2011 Jan;2(1):32-42.
49. Pan Z, Zhang X, Heflin J. DLDB2: A scalable multi-perspective semantic web repository. Web Intelligence 2008:489-495. [doi: 10.1109/WIIAT.2008.290]
50. Zhou J, Ma L, Liu Q, Zhang L, Yu Y, Pan Y. Minerva: a scalable OWL ontology storage and inference system. The Semantic Web–ASWC 2006 2006;4185:429-443. [doi: 10.1007/11836025_42]

Abbreviations

API: application programming interface
CAPS: catastrophic antiphospholipid syndrome
DBMS: database management system
EAV: Entity-Attribute-Value
EMR: electronic medical record
EPR: electronic patient record
GUI: graphical user interface
ICT: information and communication technologies
IP: Internet Protocol
IRI: internationalized resource identifiers
JSTL: JSP Standard Tag Library
MDA: model-driven architecture
MVC: model-view-controller
OMG: Object Management Group
OntoCRF: Onto Clinical Research Forms
OntoDDB-MM: ontology-driven database metamodel
OWL-DB: OWL database
OWL: Ontology Web Language
RDF: Resource Description Framework
SPARQL: Simple Protocol and RDF Query Language
SUS: System Usability Scale

Edited by G Eysenbach; submitted 14.10.13; peer-reviewed by S Rudolph, T Groza; comments to author 23.11.13; revised version received 24.02.14; accepted 28.04.14; published 01.08.14.

Please cite as:
Lozano-Rubí R, Pastor X, Lozano E
OWLing Clinical Data Repositories With the Ontology Web Language
JMIR Med Inform 2014;2(2):e14
URL: http://medinform.jmir.org/2014/2/e14/
doi:10.2196/medinform.3023
PMID:25599697

Original Paper

Incorporation of Personal Single Nucleotide Polymorphism (SNP) Data into a National Level Electronic Health Record for Disease Risk Assessment, Part 1: An Overview of Requirements

Timur Beyan[1], MD, PhD; Yeşim Aydın Son[1], MD, PhD

Informatics Institute, Department of Health Informatics, Middle East Technical University, Ankara, Turkey

Corresponding Author:
Yeşim Aydın Son, MD, PhD
Informatics Institute
Department of Health Informatics
Middle East Technical University
Üniversiteler Mahallesi Dumlupınar Bulvarı No:1
ODTÜ Enformatik Enstitüsü B-207
Ankara, 06800
Turkey
Phone: 90 312 210 7708
Fax: 90 312 210 3745
Email: yesim@metu.edu.tr

Abstract

Background: Personalized medicine approaches provide opportunities for predictive and preventive medicine. Using genomic, clinical, environmental, and behavioral data, tracking and management of individual wellness is possible. A prolific way to carry this personalized approach into routine practices can be accomplished by integrating clinical interpretations of genomic variations into electronic medical records (EMRs)/electronic health records (EHRs). Today, various central EHR infrastructures have been constituted in many countries of the world including Turkey.

Objective: The objective of this study was to concentrate on incorporating the personal single nucleotide polymorphism (SNP) data into the National Health Information System of Turkey (NHIS-T) for disease risk assessment, and evaluate the performance of various predictive models for prostate cancer cases. We present our work as a miniseries containing three parts: (1) an overview of requirements, (2) the incorporation of SNP into the NHIS-T, and (3) an evaluation of SNP incorporated NHIS-T for prostate cancer.

Methods: For the first article of this miniseries, the scientific literature is reviewed and the requirements of SNP data integration into EMRs/EHRs are extracted and presented.

Results: In the literature, basic requirements of genomic-enabled EMRs/EHRs are listed as incorporating genotype data and its clinical interpretation into EMRs/EHRs, developing accurate and accessible clinicogenomic interpretation resources (knowledge bases), interpreting and reinterpreting of variant data, and immersing of clinicogenomic information into the medical decision processes. In this section, we have analyzed these requirements under the subtitles of terminology standards, interoperability standards, clinicogenomic knowledge bases, defining clinical significance, and clinicogenomic decision support.

Conclusions: In order to integrate structured genotype and phenotype data into any system, there is a need to determine data components, terminology standards, and identifiers of clinicogenomic information. Also, we need to determine interoperability standards to share information between different information systems of stakeholders, and develop decision support capability to interpret genomic variations based on the knowledge bases via different assessment approaches.

(JMIR Med Inform 2014;2(2):e15) doi:10.2196/medinform.3169

KEYWORDS

health information systems; clinical decision support systems; disease risk model; electronic health record; epigenetics; personalized medicine; single nucleotide polymorphism

Introduction

The digital age is revolutionizing the old and historical population-based health care paradigm toward personalized medicine. Traditional medical approaches are not sufficiently predictive and preventive, as they focus on the manifestation of symptoms that often hide risk factors. Determining risk factors allows for prevention through early diagnosis, and provides new opportunities for developing personalized medicine approaches based on patient-centered, predictive, preventive, and effective health care services [1].

Genomic data and its derivatives (transcriptomes, proteomes, metabolomes, etc) are the essential elements of personalized medicine [2,3]. Every individual has almost four million variations in their own genome, when compared to the reference sequence. Genomic variations can range from single nucleotide changes to the gain or loss of whole chromosomes. Single nucleotide polymorphisms (SNPs), where a single nucleotide in the genome alters between individual or paired chromosomes, are about 90% of genomic variants, and some are already validated as important markers in the clinical practice, while others are on the way [4-6].

The rapid developments in next generation sequencing (NGS) technologies have substantially reduced both the cost and the time required to sequence the entire human genome, and it is expected that NGS-based analyses, for example, whole genome sequencing (WGS) and whole exome sequencing (WES), will be available for routine use in health care and prevention of disease by 2020 [7]. Providing genomic data to medical professionals will facilitate clinical decisions based on the individual's genome, and allow tailoring health care services to the patient's specific needs and characteristics [8]. In parallel, direct-to-consumer (DTC) genome-wide profiling tests are being developed to assess individual disease risks for many common polygenic diseases [9]. DTC genomic companies, for example, 23andMe, GenePlanet, and DNA DTC generally perform a gene-chip analysis of SNPs using deoxyribonucleic acid (DNA) extracted from saliva or serum sample [10-12].

In clinical decision processes, genomic variant data can be used for assessing disease risks, predicting susceptibility, early clinical diagnosing, following the course of the disease, targeted screening, and planning treatment regimens [3,13]. A reasonable way to carry this personalized approach into routine for medical practices would be integrating genotype data and its clinical interpretation within the electronic medical records (EMRs)/electronic health records (EHRs) [8,14].

Today, in many developed and developing countries, use of EMRs/EHRs is inevitable for health care providers for reimbursement of services, and to track the quality of the health care provided [15,16]. Recently, several EHR networks have been constituted in many countries of the world, including the National Health Information System of Turkey (NHIS-T) [17]. These EHR systems and networks have high potential for integrating genomic data in health care practices for personalized medicine.

In this work, as an initial attempt to develop a sophisticated infrastructure, we focused to incorporate the personal SNP data into NHIS-T for disease risk assessment, and evaluated the performance of various predictive models for prostate cancer cases. We presented our work as three parts: (1) a literature review for requirements, (2) the incorporation of SNP into the NHIS-T [18], and (3) an evaluation of SNP incorporated into NHIS-T for prostate cancer [19]. In this part, the scientific literature was reviewed, and the requirements were extracted regarding SNP data integrated EMRs/EHRs.

Methods

The informatics pipeline for genome sequencing can be divided into several analytical steps, for example, base calling, alignment, variant analysis, interpretation, and in all levels different file formats are generated [20-22]. Currently, tools and techniques are developed for automated and reliable analysis, but clinical interpretation of variant data is still a major problem [21].

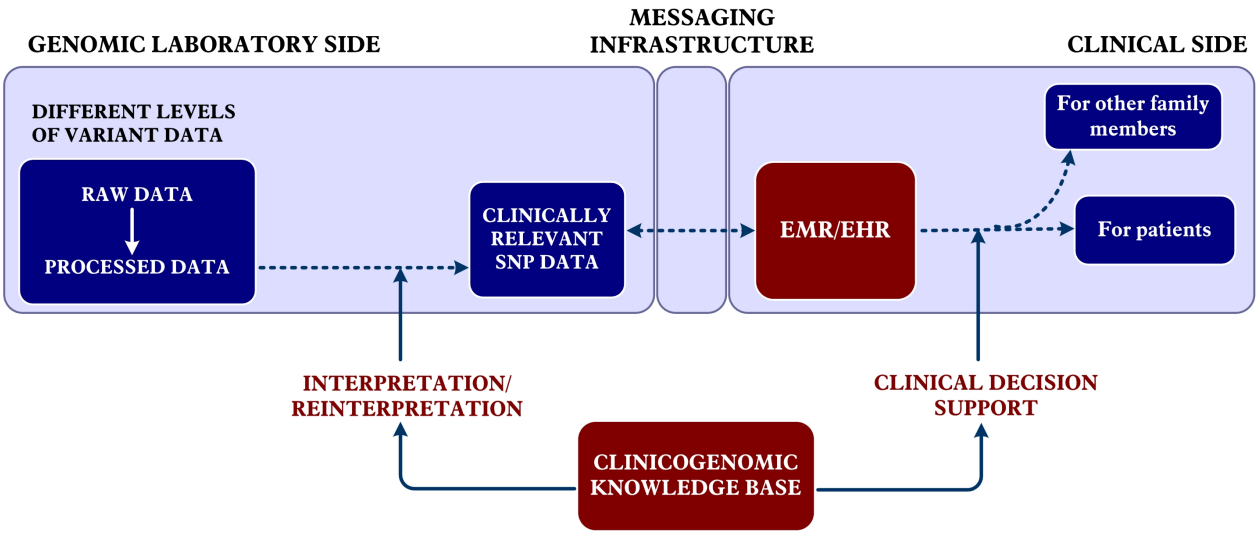
Today, most of the EMRs/EHRs are designed to store and retrieve the laboratory values and clinical findings, but do not have the ability to manage genomic data [23-25]. After WGS/WES, a file that contains a large number of variant data is acquired [26]. An entire genome sequence (the size of the haploid human genome) contains about 3 billion base pairs, and a single WGS data file is about 3 gigabytes. Storing and sharing of personal raw genomic sequences exceeds the transmission and storage capacity in many health care organizations [27]. Due to these technical limitations, raw genomic data are generally stored outside of the EMR; similar to picture archiving and communication systems for medical images, and clinical interpretation of the genomic data is preferably sent to the database of the EMR [28-30].

The initiatives of integrating a patient's genomic data into EMRs/EHRs is of a preliminary nature [31], and, until recently, only a few successful systems have been established, such as Cerner's Genomics Solutions, McKesson's Horizon Clinicals, and GeneInsight [26,32].

In the literature, basic requirements of genomic-enabled EMRs/EHRs are listed as incorporating genotype data and its clinical interpretation into EMRs/EHRs, developing accurate and accessible clinicogenomic interpretation resources (knowledge base), the interpretation and reinterpretation of variant data, and the immersion of clinicogenomic information into the medical decision processes.

Figure 1 shows, in the genome laboratory side, various levels of sequence data can be produced. Since clinicians need an actionable clinical interpretation of the variant data, it is sufficient to share clinically relevant data between the laboratory and the clinical systems. The development of a clinicogenomic knowledge base is an obligation to extract clinical meaning from the variant data. On the clinical side, it is necessary to use decision support systems due to the high number of variants. In some cases, clinicogenomic information may be useful to manage the health status of other family members and other close relatives.

Figure 1. Main components of a genome-enabled electronic medical record/electronic health record. SNP; single nucleotide polymorphisms.



Results

Terminology Standards

In order to integrate structured genotype and phenotype data into any system, the first requirement is to determine data components, terminology standards, and identifiers of clinicogenomic information, for example, genotype data and its associated clinical interpretation.

In genomic terminology, the Human Gene Nomenclature Committee standardizes identifying gene symbols, identifiers, and variant nomenclature defined by the Human Genome Variation Society [6]. Reference SNP number (rs number) and reference SNP identifier (rsID) are used to identify every single SNP entry in the Single Nucleotide Polymorphism Database (dbSNP), which is the largest database maintained by the National Center for Biotechnology Information (NCBI). The dbSNP is interconnected with many other resources, for example, Entrez Gene, GenBank, the Universal Protein Resource, the International HapMap Project, the Pharmacogenomics Knowledge Base (PharmGKB), and the AlzGene, PDGene, SzGene databases through the rsID [33]. Additionally, in many types of personal genomic file formats (eg, 23andMe, deCODEme, and Navigenics), SNPs are identified by rsID.

DNA is a double stranded stretch, and every nucleated somatic cell has 22 pairs of autosomal, and one pair of sex chromosomes. This means for autosomal chromosomes we have two versions of DNA strands inherited via maternal and paternal sex cells. Different forms or variants of a particular polymorphism are called alleles. Because different alleles may have different degrees and types of clinical impact, rsID is insufficient alone to identify the clinicogenomic significance of SNPs. To have a heterozygote allele may not change the risk for the disease, but homozygote allele of the same SNP variant may change the risk for a disease dramatically. For example, in a study, the odds ratio for rs3218536 (A;G) was 0.8 (CI 0.7-1.0), and for the rs3218536 (A;A) 0.3 (0.1-0.9) [34]. Consequently, to identify clinically relevant SNPs, we need to use a combination of rsID and allele data as the minimum requirements.

DNA has a double strand (plus and minus or forward and reverse stands respectively), and every SNP can be identified using either of the two DNA strands. In various publications, the same alleles of SNPs are defined differently based on the orientation discrepancy [35]. Due to the double-stranded structure of DNA, both approaches are correct, but it is required to declare and use a standard.

Integration of variant data and clinical relevancies bring out the issue of terminological standardization. Unfortunately, conventional health information terminologies do not successfully support the genetic diseases. There is a critical gap between the databases, which involve many terms defining the genetic diseases, and the Systematized Nomenclature of Medicine (SNOMED) [36]. In order to address the chasm between medical vocabularies and bioinformatics resources, the clinical bioinformatics ontology (CBO) was developed and implemented. The CBO is a curated semantic network trying to combine a variety of clinical vocabularies, for example, SNOMED-Clinical Term (CT), Logical Observation Identifiers Names and Codes (LOINC), and NCBI bioinformatics resources [37,38].

In addition, the International Classification of Diseases (ICD) codes, which is also implemented in Turkey, is also preferred for identifying clinical conditions, but the released versions of ICD do not fully support genomic medicine [36]. Existing ICD versions are not efficient to manage all of the levels of clinical, pathologic, and genetic heterogeneities. It is expected that these will be managed in the next version, for example, ICD-11. The ICD-11, which is scheduled for release in 2015, is expected to be interoperable with other medical terminologies such as SNOMED-CT [39]. Nevertheless, it is an unavoidable requirement to develop a new taxonomy of diseases that will be based on information commons and knowledge networks, including a combination of molecular, social, environmental, and clinical data and health outcomes [40].

As explained in the next section, in the clinicogenomic knowledge base, the assessment of both evidence quality of study and effect size of these associations are critical for the analysis of the published results for clinicogenomic associations [41-47]. Despite emerged approaches and initiatives, standardized definitions and value assignment approaches are needed to categorize and use these associations in a consistent way.

Especially for polygenic complex diseases, impact degrees of clinicogenomic association may be different according to race, ethnicity, and environmental factors [48]. The terms of "ethnicity" and "race" refer to a sociocultural construct affecting both biological and environmental factors, and we need a general standard to categorize these terms.

Various predictive models, in clinical settings, may be useful to assess personal disease risk using relevant SNPs, for example, cumulative models, polygenic risk scores, etc. On the other hand, only a small number of holistic enviro-genomic models are available. Because most of the complex diseases are progressing as the interaction of genomic and environmental factors, it seems that, more enviro-genomic models will be produced in near future. Naturally, with the increase of the number and the value of predictive clinicogenomic models, we will need standardized definition and sharing methods for these models.

Interoperability Standards

Health Level 7 (HL7) is a global organization developing health information standards. As an interoperability standard, the HL7 version 2.x (HL7 v2.x) is the most widely used all over the world. HL7 v2.x does not have not a clear information model, and contains many optional data fields. To overcome this vagueness problem, HL7 version 3 (HL7 v3) has been developed, which is based on an object oriented data model called Reference Information Model (RIM) [49]. HL7 v3 Clinical Document Architecture (CDA) is a document markup standard. A HL7 CDA document is produced to exchange information as part of the HL7 v3 standards, and aim to specify the structural and semantic aspects of clinical documents [50].

The HL7 Clinical Genomics (CG) Work Group (WG) develops standards intended to regulate interoperability issues in genomic medicine. *HL7 Version 2 Implementation Guide: Clinical Genomics; Fully LOINC-Qualified Genetic Variation Model* is based on both the *HL7 Version 2 Implementation Guide Laboratory Result Reporting to the EHR*, and the *HL7 Version 3 Genetic Variation* data model. This guide covers the reporting of the test results for sequencing and genotyping tests, and includes testing for DNA variants associated with diseases and pharmacogenomic applications [36,51,52]. *HL7 Version 2 Implementation Guide: Clinical Genomics; Fully LOINC-Qualified Genetic Variation Model* was the first example used by The Partners HealthCare Center for Personalized Genetic Medicine and the Intermountain Healthcare Clinical Genetics Institute to gather genetic test results and transmit them to a patient's EHR [51,52]. GeneInsight Suite (GeneInsight Lab, GeneInsight Clinic, and GeneInsight Network) is also a platform where clinical variant data sharing was based on HL7 standards [26,29,53,54].

The HL7 v3 genetic variation specification is based on the HL7 RIM. It uses the HL7 data types, vocabulary binding mechanisms built into the RIM and Bioinformatic Sequence Markup Language to model the sequence information. The root class in the genetic variation model is "genetic loci", which describes a set of loci, such as a haplotype, a genetic profile, and genetic testing results of multiple variations or gene expression panels. The genetic loci model uses the genetic locus as an information unit to describe each of these loci. A genetic locus is composed of one or more individual alleles, sequences, and observed sequence variations and represents a single gene or coding region. Within this model, HL7 suggests the sharing of the essential part of raw genomic data via "encapsulation", and extracting clinically relevant data via "bubble-up" based on a genomic decision support application [55].

HL7 CG-WG develops a CDA implementation guide (ie, Implementation Guide for CDA Release 2 Genetic Testing Report) to ensure the transmission of genetic testing reports using HL7 v3 RIM, and is appropriate for the level of granularity of human-readable reports [56].

Clinicogenomic Knowledge Bases

Clinicians cannot extract clinical interpretation of variants directly from the medical sources due to temporal and cognitive limitations [57,58]. So, instead of incorporating all sequence data, integration of the clinical interpretations of variant data into medical records will be more efficient for clinical decision making [54,59]. Therefore, clinically relevant variants must be selected and presented with their clinical meaning, for example, clinicogenomic associations, along with an action plan for clinicians. Since the Human Genome Project, researchers have been discovering new clinicogenomic associations continuously, and it is critical to reinterpret variants and integrate new clinical interpretations into clinical processes [26].

Clinicogenomic associations, which are acquired via studies based on a candidate gene investigation or agnostic screening of complete genome, are published in the scientific literature [41]. Some clinicogenomic knowledge bases collect, curate, interpret, and categorize these published associations between genomic variations and clinical conditions. The Cancer Genome-wide Association and Meta Analyses Database is a part of Cancer Genomic Evidence-based Medicine Knowledge Base, and provides genome-wide association studies (GWAS), research, and meta-analysis about clinicogenomic associations [60,61]. ClinVar provides reports for variations and related phenotypes with evidences [62]. AlzGene [63], PDGene [64], and SzGene [65] are resources, which include manually curated PubMed articles, using systematic methods for Alzheimer's disease, Parkinson's disease, and schizophrenia, respectively. SNPedia is a wiki resource of human genetic variation as published in peer-reviewed research [66]. PharmGKB is a knowledge source containing clinically relevant genotype-phenotype and gene-drug relationships [67].

However, many of the existing knowledge bases for the clinical interpretation of variant data have different conventions. Also, they are not error proof and are not sustainable due to funding issues [54]. Especially for polygenic complex diseases, the impact degrees of clinicogenomic association may be different

according to race, ethnicity, and environmental factors [48]. Therefore, in personalized risk assessment, it will be an ideal approach to use population specific clinicogenomic results, or at least findings from similar communities. If these are not possible, it might be conceivable to use other scientific resources with a confidence range. Experts have been advocating for the generation of centrally curated national repositories of clinically significant variants for the interpretation of an individual's genomic information, eventually [58,68]. To develop a national level clinicogenomic knowledge base is critical to consider consistency of clinicogenomic associations with the sociodemographic characteristics of citizens, and overcome the issues about sustainability.

Regarding published results of clinicogenomic associations, two major points are significant, evidence quality of study and effect size of these associations [41,42]. To measure the magnitude of impact for clinicogenomic associations, researchers usually prefer to use conventional approaches, for example, odds ratio (OR) and relative risks for case control studies and cohort studies, respectively. These values are presented with CI [43].

In GWAS, many defects and biases might be present based on study design, genotyping, or collected data quality that will affect the clinical value of results [41,44,45]. The quality of evidence is scored based on the type of study and how well the study is conducted [46], and some guidelines are proposed to calculate the evidence degree [47].

Human Genome Epidemiology Network has published the interim Venice guidelines to grade the cumulative evidence in genetic associations. This guideline is based on three criteria: (1) the amount of evidence (sample size), (2) replication of studies (determining association in different studies), and (3) protection from bias (Table 1). After the evaluation of a study, all considerations are categorized as A, B, and C, and finally, merged as a composite assessment using a semiquantitative index as strong, moderate, and weak epidemiological credibility for genetic associations [47].

Table 1. Venice interim guideline criteria for assessment of cumulative evidence on genetic associations [47].

Venice interim guideline criteria	Categories
Amount of evidence	Category A, sample size >1000
	Category B, sample size >100 and <1000,
	Category C, sample size <100
	(total number in cases and controls assuming 1:1 ratio)
Extent of replication	Category A, extensive replication including at least one well conducted meta-analysis with little between-study inconsistency.
	Category B, well conducted meta-analysis with some methodological limitations or moderate between-study inconsistency.
	Category C, no association; no independent replication; failed replication; scattered studies; flawed meta-analysis; or large inconsistency.
Protection from bias	Category A, bias, if at all present, could affect the magnitude, but probably not the presence of the association.
	Category B, no obvious bias that may affect the presence of the association, but there is considerable missing information on the generation of evidence.
	Category C, considerable potential for, or demonstrable bias, which can affect even the presence or absence of the association.

Defining Clinical Significance

Today, Venice criteria are used to assess genomic association studies in several controlled and structured knowledge bases, for example, Alz-Gene, PD-Gene, and SZ-Gene [63-65]. For the importance of clinicogenomic association, some of the knowledge sources include additional data fields that define the magnitude of clinical effects and strength of the relationship between variants and diseases. In ClinVar, clinical significance is defined as a combination of impact and clinical function (eg, benign, pathogenic, protective, drug response, etc), and evidence for clinical significance is categorized regarding study count and type, such as in vitro studies, animal models, etc [62]. In the PharmGKB, a systematic categorization for evidence quality of clinicogenomic associations is extracted depending on methods and results of references [67], but impact value is not emphasized as a parameter. In SNPedia, magnitude is constructed as a subjective measure of interest for magnitude of impact and repute (good, bad) for quality of evidence, but these concepts are not well established. In GET-Evidence, clinicogenomic references are categorized according to their evidence degree (high, moderate, or low), and clinical significance (high, medium, or low) is used to produce impact score [69].

Clinicogenomic Decision Support

The volume of variation data integrated into clinical practice exceeds the boundaries of unsupported human cognition and interpretive capacity. Additionally, the rapidly growing literature on clinicogenomic associations makes it more complicated to stay current for even experienced professionals [29]. Also, it is not reasonable to expect the interpretation of all clinicogenomic data by the limited number of genetics experts; we need more automated solutions to overcome these obstacles [70]. With the growing data load in the genomic era, in order to make informed decisions in a timely manner, the health care systems need to shift from expert-based practice to systems-supported practice [71].

Although there is a limited number of counter examples, in general, the clinical effect of a single SNP is minor (OR <2.00) [72,73]. Nevertheless, listing of clinicogenomic associations and their effects may be useful to report a limited number of independent associations. This is especially true for disease-associated SNPs with strong impact and strong evidence; users can share these one by one. At this point, using carefully chosen graphics and visualization techniques will be an efficient way of doing so. Various DTC genomic companies report personal genomic risk for various clinical conditions using graphics containing personal estimations [74].

Although the simplest way of reporting SNP variations is displaying these numerous variations in laboratory reports, it is clear that clinicians cannot interpret or evaluate this information stack. Modest value of clinicogenomic associations does not mean negligible, and some researchers try to develop polygenic risk models or panels assigning values for various SNP alleles, and calculate the total risk of disease for more effective risk prediction [75]. In the literature, several cumulative prediction models have been proposed, but most of these are criticized regarding comprehensive evaluation, especially for clinical utility [76].

SNPs could be used to produce a "genomic profile" for disease risk prediction, testing hundreds of thousands of loci across the personal genome. Today, most of the SNP-based risk assessment models have limited predictive utility and discriminative accuracy because most of the disease associated SNPs have small impacts [77,78]. It has been suggested that genomic risk scores based on large numbers of SNPs could explain more about the heritability than models based on a small number and rigorously validated SNPs. But there is a requirement to process large datasets to build such discriminative risk assessment models [79,80].

The genetic architecture of a disease refers to the number, effect size, genetic mode of action (additive, dominant, and/or epistatic), and allelic frequencies of the genetic polymorphisms. The prediction of genetic risk depends on the underlying genetic architecture. Indeed, the SNPs do not have to be the causative mutations. They just need to be in high linkage disequilibrium with the causative mutations so that there is a consistent association between the SNP and disease risk [81].

Different types of polygenic prediction models have been developed to combine the impact of disease associated SNP data, for example, count method, log odds method, multiplicative model, etc. The count method is the calculation of the total count of independent genomic risk alleles. The log odds method sums together the natural logarithm of the allelic OR for each risk allele [78]. DTC testing companies typically employ a multiplicative model to calculate lifetime risk in the absence of an established method for combining SNP risk estimates, for example, multiplication of ORs of each genotype and average population risk [82].

There are various cumulative models combining the impact of several clinicogenomic associations using arithmetic operators. In recessive models, only homozygote alleles are involved in the models, but in dominant models heterozygote SNPs are also a part of the cumulative models. Both in dominant and recessive models, the values of risk SNPs are accepted as one unit of impact. Models involving alterations of SNPs' impact value regarding homozygote and heterozygote alleles are defined as an additive model [35,43].

Some of the models involve additional criteria, for example, family history [83]. But structured family history is not a mandatory part of EHR, and because of its dynamic characteristics, it is reasonable to collect and trace it at each visit from patients. It is clear that, similar to clinicogenomic associations, collection and reinterpretation of family history is critical to capture effective results with this type of predictive models.

Actually, genomic and environmental factors are involved in various degrees with the molecular etiology of diseases. In monogenetic diseases (eg, Huntington's disease, phenylketonuria, hereditary cancer forms, etc), single gene mutations are predominantly the main cause of diseases. The genetic origins of the complex multifactorial diseases are much more complicated than the monogenetic diseases, which are a result of the complicated interactions between genetic and environmental causes [84].

Genomic information has lifelong value and one's genomic findings can reveal others within families [23]. If a patient is found to have a disease associated variant, possibly other blood relatives would carry the similar risk, and the patient's health care provider could utilize this new clinical information [26]. This is especially important, not only because of the medical perspective, but also for security and privacy issues.

Discussion

In this part of the miniseries, we have reviewed the scientific literature to extract the requirements for SNP data integrated into EMRs/EHRs.

In order to integrate structured genotype and phenotype data into any system, the first requirement is to determine data components, terminology standards, and identifiers of clinicogenomic information, for example, genotype data and its associated clinical interpretation. Also, we need interoperability standards such as HL7 v2 or v3 to share information between stakeholders.

Because of the huge amount of clinically relevant genomic data and fast translation of this information to a clinical domain, we need clinical decision support capability. To ensure this capability, we also need a continuously updated accredited and structured knowledge base, and assessment approaches to interpret these genomic variations.

In the next part of the miniseries, we will present our study to extend capabilities of NHIS-T to handle SNP data, and its clinical interpretation to assess personal disease risk, and propose possible solutions regarding these requirements.

Conflicts of Interest

None declared.

References

1. Downing GJ. Key aspects of health system change on the path to personalized medicine. Transl Res 2009 Dec;154(6):272-276. [doi: 10.1016/j.trsl.2009.09.003] [Medline: 19931192]
2. Arnold GL, Vockley J. Thoroughly modern medicine. Mol Genet Metab 2011;104(1-2):1-2. [doi: 10.1016/j.ymgme.2011.07.011] [Medline: 21807540]
3. Ginsburg GS, Willard HF. Genomic and personalized medicine: Foundations and applications. Transl Res 2009 Dec;154(6):277-287. [doi: 10.1016/j.trsl.2009.09.005] [Medline: 19931193]
4. Barnes MH. Genetic variation analysis for biomedical researchers: A primer. In: Genetic variation: Methods and protocols (methods in molecular biology). USA: Humana Press; 2010.
5. Drmanac R. Medicine. The ultimate genetic test. Science 2012 Jun 1;336(6085):1110-1112. [doi: 10.1126/science.1221037] [Medline: 22654043]
6. Poo DC, Cai S, Mah JT. UASIS: Universal automatic SNP identification system. BMC Genomics 2011 Nov 30;12 Suppl 3:S9 [FREE Full text] [doi: 10.1186/1471-2164-12-S3-S9] [Medline: 22369494]
7. Berg JS, Khoury MJ, Evans JP. Deploying whole genome sequencing in clinical practice and public health: Meeting the challenge one bin at a time. Genet Med 2011 Jun;13(6):499-504. [doi: 10.1097/GIM.0b013e318220aaba] [Medline: 21558861]
8. Scheuner MT, de Vries H, Kim B, Meili RC, Olmstead SH, Teleki S. Are electronic health records ready for genomic medicine? Genet Med 2009 Jul;11(7):510-517. [doi: 10.1097/GIM.0b013e3181a53331] [Medline: 19478682]
9. Bloss CS, Schork NJ, Topol EJ. Effect of direct-to-consumer genomewide profiling to assess disease risk. N Engl J Med 2011 Feb 10;364(6):524-534 [FREE Full text] [doi: 10.1056/NEJMoa1011893] [Medline: 21226570]
10. Helgason A, Stefánsson K. The past, present, and future of direct-to-consumer genetic tests. Dialogues Clin Neurosci 2010;12(1):61-68 [FREE Full text] [Medline: 20373667]
11. Chua EW, Kennedy MA. Current state and future prospects of direct-to-consumer pharmacogenetics. Front Pharmacol 2012;3:152 [FREE Full text] [doi: 10.3389/fphar.2012.00152] [Medline: 22934000]
12. Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Becich MJ. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. J Pathol Inform 2012;3:40 [FREE Full text] [doi: 10.4103/2153-3539.103013] [Medline: 23248761]
13. Chan IS, Ginsburg GS. Personalized medicine: Progress and promise. Annu Rev Genomics Hum Genet 2011;12:217-244. [doi: 10.1146/annurev-genom-082410-101446] [Medline: 21721939]
14. Belmont J, McGuire AL. The futility of genomic counseling: Essential role of electronic health records. Genome Med 2009;1(5):48 [FREE Full text] [doi: 10.1186/gm48] [Medline: 19439060]
15. Garets D, Davis M. Electronic medical records vs electronic health records: Yes, there is a difference. 2006 URL: http://www.himssanalytics.org/docs/WP_EMR_EHR.pdf [accessed 2013-12-03] [WebCite Cache ID 6Lb0ZAXxG]
16. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. Int J Med Inform 2008 May;77(5):291-304. [doi: 10.1016/j.ijmedinf.2007.09.001] [Medline: 17951106]
17. Healthcare Information and Management Systems Society (HIMSS) Global Enterprise Task Force. Electronic health records: A global perspective, part 1. 2010 URL: http://www.himss.org/files/HIMSSorg/content/files/Globalpt1-edited%20final.pdf [accessed 2013-12-03] [WebCite Cache ID 6Lb0VPAh0]
18. Beyan T, Aydın Son Y. Incorporation of personal single nucleotide polymorphism (SNP) data into a national level electronic health record for disease risk assessment, part 2: The incorporation of SNP into the National Health Information System of Turkey. JMIR Med Inform 2014 (forthcoming). [doi: 10.2196/medinform.3555]
19. Beyan T, Aydın Son Y. Incorporation of personal single nucleotide polymorphism (SNP) data into a national level electronic health record for disease risk assessment, part 3: Evaluation for prostate cancer risk assessment. JMIR Med Inform 2014 (forthcoming). [doi: 10.2196/medinform.3560]
20. Röhm U, Blakeley JA. Data management for high-throughput genomics. 2009 Presented at: CIDR , Fourth Biennial Conference on Innovative Data Systems Research; Asilomar, CA; January , 2009; Asilomar, CA, US p. 4-7 URL: http://www-db.cs.wisc.edu/cidr/cidr2009/Paper_31.pdf
21. Wright C, Burton H, Hall A, Moorthie S, Pokorska-Bocci A, Sagoo G, et al. Next steps in the sequence: The implications of whole genome sequencing for health in the UK. Cambridge, UK: PHG Foundation; 2011.
22. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 2011 Jun;12(6):443-451 [FREE Full text] [doi: 10.1038/nrg2986] [Medline: 21587300]
23. Hoffman MA. The genome-enabled electronic medical record. J Biomed Inform 2007 Feb;40(1):44-46 [FREE Full text] [doi: 10.1016/j.jbi.2006.02.010] [Medline: 16616698]

24. Sethi P, Theodos K. Translational bioinformatics and healthcare informatics: Computational and ethical challenges. Perspect Health Inf Manag 2009;6:1h [FREE Full text] [Medline: 20169020]
25. Jacob HJ, Abrams K, Bick DP, Brodie K, Dimmock DP, Farrell M, et al. Genomics in clinical practice: Lessons from the front lines. Sci Transl Med 2013 Jul 17;5(194):194-195. [doi: 10.1126/scitranslmed.3006468] [Medline: 23863829]
26. Aronson SJ, Clark EH, Varugheese M, Baxter S, Babb LJ, Rehm HL. Communicating new knowledge on previously reported genetic variants. Genet Med 2012 Apr 5 [FREE Full text] [doi: 10.1038/gim.2012.19] [Medline: 22481129]
27. Kahn SD. On the future of genomic data. Science 2011 Feb 11;331(6018):728-729. [doi: 10.1126/science.1197891] [Medline: 21311016]
28. Starren J, Bottinger E, Dente M, Wood G, Hoffman J. AMIA Summit on Translational Bioinformatics. 2012. Crossing the omic chasm: Integrating omic data into the EHR URL: http://knowledge.amia.org/amia-55142-tbi2012a-1.649213/t-003-1.649838/f-001-1.649839/a-017-1.649846/an-017-1.649847?qr=1 [accessed 2014-07-18] [WebCite Cache ID 6R4C6yAtB]
29. Masys DR, Jarvik GP, Abernethy NF, Anderson NR, Papanicolaou GJ, Paltoo DN, et al. Technical desiderata for the integration of genomic data into electronic health records. J Biomed Inform 2012 Jun;45(3):419-422 [FREE Full text] [doi: 10.1016/j.jbi.2011.12.005] [Medline: 22223081]
30. Green RC, Rehm HL, Kohane IS. Clinical genome sequencing. In: Ginsburg GS, Willard HF, editors. Genomic and personalized medicine, 2nd ed. Amsterdam, Netherland: Academic Press; 2013:102-122.
31. Jing X, Kay S, Marley T, Hardiker NR, Cimino JJ. Incorporating personalized gene sequence variants, molecular genetics knowledge, and health knowledge into an EHR prototype based on the Continuity of Care Record standard. J Biomed Inform 2012 Feb;45(1):82-92 [FREE Full text] [doi: 10.1016/j.jbi.2011.09.001] [Medline: 21946299]
32. Gerhard GS, Carey DJ, Steele Jr GD. Electronic health records in genomic medicine. In: Genomic and personalized medicine, 2nd ed. Amsterdam, Netherland: Academic Press; 2013:287-294.
33. Thomas PE, Klinger R, Furlong LI, Hofmann-Apitius M, Friedrich CM. Challenges in the association of human single nucleotide polymorphism mentions with unique database identifiers. BMC Bioinformatics 2011;12 Suppl 4:S4 [FREE Full text] [doi: 10.1186/1471-2105-12-S4-S4] [Medline: 21992066]
34. Auranen A, Song H, Waterfall C, Dicioccio RA, Kuschel B, Kjaer SK, et al. Polymorphisms in DNA repair genes and epithelial ovarian cancer risk. Int J Cancer 2005 Nov 20;117(4):611-618. [doi: 10.1002/ijc.21047] [Medline: 15924337]
35. Attia J, Ioannidis JP, Thakkinstian A, McEvoy M, Scott RJ, Minelli C, et al. How to use an article about genetic association: A: Background concepts. JAMA 2009 Jan 7;301(1):74-81. [doi: 10.1001/jama.2008.901] [Medline: 19126812]
36. Ullman-Cullere MH, Mathew JP. Emerging landscape of genomics in the electronic health record for personalized medicine. Hum Mutat 2011 May;32(5):512-516. [doi: 10.1002/humu.21456] [Medline: 21309042]
37. Hoffman M, Arnoldi C, Chuang I. The clinical bioinformatics ontology: A curated semantic network utilizing RefSeq information. Pac Symp Biocomput 2005:139-150. [Medline: 15759621]
38. Hoffman MA, Williams MS. Electronic medical records and personalized medicine. Hum Genet 2011 Jul;130(1):33-39. [doi: 10.1007/s00439-011-0992-y] [Medline: 21519832]
39. Zafar A, Ezat WP S. Development of ICD 11: Changes and challenges. BMC Health Serv Res 2012;12(Suppl 1):I8. [doi: 10.1186/1472-6963-12-S1-I8]
40. Committee on a Framework for Development a New Taxonomy of Disease, National Research Council. Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease. USA: National Academies Press; 2011.
41. Attia J, Ioannidis JP, Thakkinstian A, McEvoy M, Scott RJ, Minelli C, et al. How to use an article about genetic association: B: Are the results of the study valid? JAMA 2009 Jan 14;301(2):191-197. [doi: 10.1001/jama.2008.946] [Medline: 19141767]
42. Van Allen EM, Wagle N, Levy MA. Clinical analysis and interpretation of cancer genome data. J Clin Oncol 2013 May 20;31(15):1825-1833. [doi: 10.1200/JCO.2013.48.7215] [Medline: 23589549]
43. Attia J, Ioannidis JP, Thakkinstian A, McEvoy M, Scott RJ, Minelli C, et al. How to use an article about genetic association: C: What are the results and will they help me in caring for my patients? JAMA 2009 Jan 21;301(3):304-308. [doi: 10.1001/jama.2008.993] [Medline: 19155457]
44. Pearson TA, Manolio TA. How to interpret a genome-wide association study. JAMA 2008 Mar 19;299(11):1335-1344. [doi: 10.1001/jama.299.11.1335] [Medline: 18349094]
45. Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, STrengthening the REporting of Genetic Association Studies. Strengthening the reporting of genetic association studies (STREGA): An extension of the STROBE statement. PLoS Med 2009 Feb 3;6(2):e22 [FREE Full text] [doi: 10.1371/journal.pmed.1000022] [Medline: 19192942]
46. Riegelman R. Public health 101: Healthy people - healthy populations (essential public health). USA: Jones & Bartlett Publishers; 2010.
47. Ioannidis JP, Boffetta P, Little J, O'Brien TR, Uitterlinden AG, Vineis P, et al. Assessment of cumulative evidence on genetic associations: Interim guidelines. Int J Epidemiol 2008 Feb;37(1):120-132 [FREE Full text] [doi: 10.1093/ije/dym159] [Medline: 17898028]
48. Stepanov VA. Genomes, populations and diseases: Ethnic genomics and personalized medicine. Acta Naturae 2010 Oct;2(4):15-30 [FREE Full text] [Medline: 22649660]

49. Benson T. Principles of health interoperability HL7 and SNOMED, 2nd ed. New York, US: Springer; 2012:121-141.
50. Boone KW. The CDA™ book. London: Springer-Verlag London Limited; 2011:17-20.
51. Shabo A, Ullman-Cullere M, Pochon P, Huff S, Wood G, McDonald C, et al. HL7 version 2 implementation guide: Clinical genomics; fully loinc-qualified genetic variation model, release 1 (1st informative ballot), HL7 version 2. 2009. URL: http://wiki.hl7.org/images/2/24/V2_CG_LOINCGENVAR_R1_I2_2009MAY.pdf [accessed 2013-11-29] [WebCite Cache ID 6LUSIAYIi]
52. Ribick A. Health level seven clinical genomics version 2 messaging standard implementation guide successfully transmits genomic data electronically. 2010 URL: http://www.hl7.org/documentcenter/public_temp_63D90767-1C23-BA17-0C09452387FC1279/pressreleases/hl7_press_20100119.pdf [accessed 2013-11-29] [WebCite Cache ID 6LURs9kk4]
53. Aronson SJ, Clark EH, Babb LJ, Baxter S, Farwell LM, Funke BH, et al. The GeneInsight Suite: A platform to support laboratory and provider use of DNA-based genetic testing. Hum Mutat 2011 May;32(5):532-536 [FREE Full text] [doi: 10.1002/humu.21470] [Medline: 21432942]
54. Health ROTGRF, Medicine IO. In: Olson S, editor. Integrating large-scale genomic information into clinical practice: Workshop summary. USA: National Academies Press; 2012.
55. Shabo A. Clinical genomics data standards for pharmacogenetics and pharmacogenomics. Pharmacogenomics 2006 Mar;7(2):247-253. [doi: 10.2217/14622416.7.2.247] [Medline: 16515405]
56. Shabo A, Ullman-Cullere MS. Implementation guide for CDA release 2 genetic testing report (GTR) (universal realm) draft standard for trial use September 2012 (developer documentation). 2012. URL: http://www.hl7.org/documentcenter/public_temp_5765BAE6-1C23-BA17-0C2BC20D21AE785C/wg/clingenomics/docs/Genetic%20Testing%20Report%20(GTR)%20-%202012.10.30.pdf [accessed 2014-04-02] [WebCite Cache ID 6OX7gPgEb]
57. Oetting WS. Clinical genetics & human genome variation: The 2008 Human Genome Variation Society scientific meeting. Hum Mutat 2009 May;30(5):852-856. [doi: 10.1002/humu.20987] [Medline: 19260058]
58. Starren J, Williams MS, Bottinger EP. Crossing the omic chasm: A time for omic ancillary systems. JAMA 2013 Mar 27;309(12):1237-1238 [FREE Full text] [doi: 10.1001/jama.2013.1579] [Medline: 23494000]
59. Marian AJ. Medical DNA sequencing. Curr Opin Cardiol 2011 May;26(3):175-180 [FREE Full text] [doi: 10.1097/HCO.0b013e3283459857] [Medline: 21415728]
60. Schully SD, Yu W, McCallum V, Benedicto CB, Dong LM, Wulf A, et al. Cancer GAMAdb: Database of cancer genetic associations from meta-analyses and genome-wide association studies. Eur J Hum Genet 2011 Aug;19(8):928-930 [FREE Full text] [doi: 10.1038/ejhg.2011.53] [Medline: 21487441]
61. Centers for Disease Control and Prevention. Cancer GAMAdb, cancer genomic evidence-based medicine knowledge base. 2013 URL: http://www.hugenavigator.net/CancerGEMKB/caIntegratorStartPage.do [accessed 2013-11-29] [WebCite Cache ID 6LUQ8RWk4]
62. National Center for Biotechnology Information. ClinVar. 2013. URL: http://www.ncbi.nlm.nih.gov/clinvar/ [accessed 2013-11-29] [WebCite Cache ID 6LUQcwkqC]
63. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. Systematic meta-analyses of Alzheimer disease genetic association studies: The AlzGene database. Nat Genet 2007 Jan;39(1):17-23. [doi: 10.1038/ng1934] [Medline: 17192785]
64. Lill CM, Roehr JT, McQueen MB, Kavvoura FK, Bagade S, Schjeide BM, 23andMe Genetic Epidemiology of Parkinson's Disease Consortium, International Parkinson's Disease Genomics Consortium, Parkinson's Disease GWAS Consortium, et al. Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: The PDGene database. PLoS Genet 2012;8(3):e1002548 [FREE Full text] [doi: 10.1371/journal.pgen.1002548] [Medline: 22438815]
65. Allen NC, Bagade S, McQueen MB, Ioannidis JP, Kavvoura FK, Khoury MJ, et al. Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: The SzGene database. Nat Genet 2008 Jul;40(7):827-834. [doi: 10.1038/ng.171] [Medline: 18583979]
66. Cariaso M, Lennon G. SNPedia: A wiki supporting personal genome annotation, interpretation and analysis. Nucleic Acids Res 2012 Jan;40(Database issue):D1308-D1312 [FREE Full text] [doi: 10.1093/nar/gkr798] [Medline: 22140107]
67. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther 2012 Oct;92(4):414-417 [FREE Full text] [doi: 10.1038/clpt.2012.96] [Medline: 22992668]
68. Kawamoto K, Lobach DF, Willard HF, Ginsburg GS. A national clinical decision support infrastructure to enable the widespread and consistent practice of genomic and personalized medicine. BMC Med Inform Decis Mak 2009;9:17 [FREE Full text] [doi: 10.1186/1472-6947-9-17] [Medline: 19309514]
69. Ball MP, Thakuria JV, Zaranek AW, Clegg T, Rosenbaum AM, Wu X, et al. A public resource facilitating clinical use of genomes. Proc Natl Acad Sci U S A 2012 Jul 24;109(30):11920-11927 [FREE Full text] [doi: 10.1073/pnas.1201904109] [Medline: 22797899]
70. Welch BM, Kawamoto K. Clinical decision support for genetically guided personalized medicine: A systematic review. J Am Med Inform Assoc 2013;20(2):388-400 [FREE Full text] [doi: 10.1136/amiajnl-2012-000892] [Medline: 22922173]
71. McClellan MB, McGinnis JM, Nabel EG, Olsen LM, Institute of Medicine. Evidence-based medicine and the changing nature of health care: 2007 IOM annual meeting summary. Washington, DC: The National Academies Press; 2008.

72. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 2011 Feb;187(2):367-383 [FREE Full text] [doi: 10.1534/genetics.110.120907] [Medline: 21115973]
73. Kalf RR, Mihaescu R, Kundu S, de Knijff P, Green RC, Janssens AC. Variations in predicted risks in personal genome testing for common complex diseases. Genet Med 2014 Jan;16(1):85-91 [FREE Full text] [doi: 10.1038/gim.2013.80] [Medline: 23807614]
74. Lautenbach DM, Christensen KD, Sparks JA, Green RC. Communicating genetic risk information for common disorders in the era of genomic medicine. Annu Rev Genomics Hum Genet 2013;14:491-513. [doi: 10.1146/annurev-genom-092010-110722] [Medline: 24003856]
75. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med 2010 Jul 8;363(2):166-176. [doi: 10.1056/NEJMra0905980] [Medline: 20647212]
76. Little J, Wilson B, Carter R, Walker K, Santaguida P, Tomiak E, et al. Multigene panels in prostate cancer risk assessment, evidence report No. 209. Rockville, MD: AHRQ Publication No.12-E020-EF; 2012 Jul. URL: http://www.effectivehealthcare.ahrq.gov/ehc/products/388/1171/EvidenceReport209_multigenepanels_FinalReport_20120629.pdf [WebCite Cache ID 6RHGsyhC7]
77. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res 2007 Oct;17(10):1520-1528 [FREE Full text] [doi: 10.1101/gr.6665407] [Medline: 17785532]
78. Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet 2009 Sep 15;18(18):3525-3531 [FREE Full text] [doi: 10.1093/hmg/ddp295] [Medline: 19553258]
79. Jostins L, Barrett JC. Genetic risk prediction in complex disease. Hum Mol Genet 2011 Oct 15;20(R2):R182-R188 [FREE Full text] [doi: 10.1093/hmg/ddr378] [Medline: 21873261]
80. Wu J, Pfeiffer RM, Gail MH. Strategies for developing prediction models from genome-wide association studies. Genet Epidemiol 2013 Dec;37(8):768-777. [doi: 10.1002/gepi.21762] [Medline: 24166696]
81. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk of complex disease. Curr Opin Genet Dev 2008 Jun;18(3):257-263. [doi: 10.1016/j.gde.2008.07.006] [Medline: 18682292]
82. Nusbaum R, Leventhal KG, Hooker GW, Peshkin BN, Butrick M, Salehizadeh Y, et al. Translational genomic research: Protocol development and initial outcomes following SNP testing for colon cancer risk. Transl Behav Med 2013 Mar 1;3(1):17-29 [FREE Full text] [doi: 10.1007/s13142-012-0149-0] [Medline: 23565131]
83. Zheng SL, Sun J, Wiklund F, Smith S, Stattin P, Li G, et al. Cumulative association of five genetic variants with prostate cancer. N Engl J Med 2008 Feb 28;358(9):910-919. [doi: 10.1056/NEJMoa075819] [Medline: 18199855]
84. Janssens AC, van Duijn CM. Genome-based prediction of common diseases: Advances and prospects. Hum Mol Genet 2008 Oct 15;17(R2):R166-R173 [FREE Full text] [doi: 10.1093/hmg/ddn250] [Medline: 18852206]

Abbreviations

CBO: clinical bioinformatics ontology
CDA: Clinical Document Architecture
dbSNP: Single Nucleotide Polymorphism Database
DNA: deoxyribonucleic acid
DTC: direct-to-consumer
EHR: electronic health record
EMR: electronic medical record
GWAS: genome-wide association studies
HL7: Health Level 7
HL7 v2.x: HL7 version 2.x
HL7 v3: HL7 version 3
HL7 CG: Health Level 7 Clinical Genomic
ICD: International Classification of Diseases
LOINC: Logical Observation Identifiers Names and Codes
NCBI: National Center for Biotechnology Information
NGS: next generation sequencing
NHIS-T: National Health Information System of Turkey
OR: odds ratio
PharmGKB: Pharmacogenomics Knowledge Base
RIM: Reference Information Model
rsID: reference SNP identifier
rs number: reference SNP number
SNOMED: Systematized Nomenclature of Medicine
SNOMED-CT: Systematized Nomenclature of Medicine Clinical Term

SNP: single nucleotide polymorphism
WES: whole exome sequencing
WG: Work Group
WGS: whole genome sequencing

Edited by G Eysenbach; submitted 08.12.13; peer-reviewed by W Hammond, A James; comments to author 31.12.13; revised version received 25.05.14; accepted 02.07.14; published 24.07.14.

Please cite as:
Beyan T, Aydın Son Y
Incorporation of Personal Single Nucleotide Polymorphism (SNP) Data into a National Level Electronic Health Record for Disease Risk Assessment, Part 1: An Overview of Requirements
JMIR Med Inform 2014;2(2):e15
URL: http://medinform.jmir.org/2014/2/e15/
doi:10.2196/medinform.3169
PMID:25599712

Original Paper

Incorporation of Personal Single Nucleotide Polymorphism (SNP) Data into a National Level Electronic Health Record for Disease Risk Assessment, Part 2: The Incorporation of SNP into the National Health Information System of Turkey

Timur Beyan[1], MD, PhD; Yeşim Aydın Son[1], MD, PhD

Informatics Institute, Department of Health Informatics, Middle East Technical University, Ankara, Turkey

Corresponding Author:
Yeşim Aydın Son, MD, PhD
Informatics Institute
Department of Health Informatics
Middle East Technical University
Üniversiteler Mahallesi Dumlupınar Bulvarı No:1
ODTÜ Enformatik Enstitüsü B-207, Çankaya
Ankara, 06800
Turkey
Phone: 90 3122107708
Fax: 90 3122103745
Email: yesim@metu.edu.tr

Abstract

Background: A personalized medicine approach provides opportunities for predictive and preventive medicine. Using genomic, clinical, environmental, and behavioral data, the tracking and management of individual wellness is possible. A prolific way to carry this personalized approach into routine practices can be accomplished by integrating clinical interpretations of genomic variations into electronic medical record (EMR)s/electronic health record (EHR)s systems. Today, various central EHR infrastructures have been constituted in many countries of the world, including Turkey.

Objective: As an initial attempt to develop a sophisticated infrastructure, we have concentrated on incorporating the personal single nucleotide polymorphism (SNP) data into the National Health Information System of Turkey (NHIS-T) for disease risk assessment, and evaluated the performance of various predictive models for prostate cancer cases. We present our work as a miniseries containing three parts: (1) an overview of requirements, (2) the incorporation of SNP into the NHIS-T, and (3) an evaluation of SNP data incorporated into the NHIS-T for prostate cancer.

Methods: For the second article of this miniseries, we have analyzed the existing NHIS-T and proposed the possible extensional architectures. In light of the literature survey and characteristics of NHIS-T, we have proposed and argued opportunities and obstacles for a SNP incorporated NHIS-T. A prototype with complementary capabilities (knowledge base and end-user applications) for these architectures has been designed and developed.

Results: In the proposed architectures, the clinically relevant personal SNP (CR-SNP) and clinicogenomic associations are shared between central repositories and end-users via the NHIS-T infrastructure. To produce these files, we need to develop a national level clinicogenomic knowledge base. Regarding clinicogenomic decision support, we planned to complete interpretation of these associations on the end-user applications. This approach gives us the flexibility to add/update envirobehavioral parameters and family health history that will be monitored or collected by end users.

Conclusions: Our results emphasized that even though the existing NHIS-T messaging infrastructure supports the integration of SNP data and clinicogenomic association, it is critical to develop a national level, accredited knowledge base and better end-user systems for the interpretation of genomic, clinical, and envirobehavioral parameters.

(JMIR Med Inform 2014;2(2):e17) doi:10.2196/medinform.3555

KEYWORDS

health information systems; clinical decision support systems; disease risk model; electronic health record; epigenetics; personalized medicine; single nucleotide polymorphism

Introduction

In clinical decision processes, genomic variant data can be used for assessing disease risks, predicting susceptibility, providing targeted screening and early diagnosis, and for planning treatment regimens [1,2]. A reasonable way to implement this personalized approach into a routine for medical practices would be to integrate genotype data and its clinical interpretation into the electronic medical record (EMR)/electronic health record (EHR) systems [3,4].

Today, in many developed and developing countries, the use of EMRs/EHRs is essential for health care providers for reimbursement of services and to track the quality of the health care provided [5,6]. Recently, several EHR networks have been established in many countries of the world, including the National Health Information System of Turkey (NHIS-T) [7]. These EHR systems and networks have high potential for integrating genomic data in health care practices for personalized medicine.

The aim of this miniseries is to present how the incorporation of personal single nucleotide polymorphism (SNP) data into the NHIS-T would make disease risk assessment possible, and to evaluate the performance of various predictive models for a specific medical condition (eg, prostate cancer). The requirements of SNP data integrated with EMRs/EHRs from scientific literature have been reviewed in the previous part of this miniseries [8], and here we will focus on extending the capabilities of the NHIS-T via incorporating SNP and clinicogenomic data for disease risk assessment. In the final part of the miniseries, we will evaluate the proposed complementary capabilities with real data from prostate cancer cases [9].

Methods

In this part of the miniseries, we studied existing NHIS-T regarding architecture, standards, and terminologies. We explained data elements, minimum health datasets, transmission sets, and the National Health Data Dictionary and their roles to produce a conceptual EHR design. Then, we clarified the messaging infrastructure and serialization approach of NHIS-T based on Health Level 7 (HL7) Clinical Document Architecture (CDA) standard.

After that, we argued the possible architectural extensions with complementary capabilities for a SNP data incorporated NHIS-T in the light of literature review and characteristics of NHIS-T [8].

After the presentation of a general use case for our approach, we put forward the design and development efforts for the complementary components, namely, knowledge base (Clinicogenomic Knowledge Base, ClinGenKB) and end-user application (Clinicogenomic Web Application, ClinGenWeb). In this phase, we have constituted the standardized definition tables for clinicogenomic associations and predictive models for these complementary capabilities.

Through analysis of the disease risk approaches from literature for prostate cancer, we have extracted possible clinicogenomic association types for assessment and reporting. At the end, we have generated a comparative table to determine the requirements to produce a standard representation for all types of clinicogenomic associations.

In addition, to interpret clinicogenomic associations at the end-user side (ClinGenWeb) using various predictive models, we designed a standardized model definition table. In both definition tables, we determined terminology standards for data elements.

In the final phase, to develop the ClinGenKB we used BioXM Knowledge Management Environment (BioXM), which is a distributed software platform providing a central inventory of information and knowledge [10]. Also, we have developed a practical reporting approach and demonstrated it using Zoho Reports as a prototype system (namely, ClinGenWeb) for the client side [11].

Results

Analysis of National Health Information System of Turkey

Overview

NHIS-T is a national level health information infrastructure that has a centralized service-oriented architecture in order to produce and share medical records among stakeholders [12,13].

Every care provider organization in Turkey has to collect patient/medical records in its EMR systems, and send some predefined structured medical data to the central Republic of Turkey's Ministry of Health (MoH) databases. The architecture of local EMR systems varies because of the existence of different vendors in the market, and generally most of the clinical data is collected as narrative texts. But, because it is mandatory to conform to the NHIS-T standards while sending predefined datasets to the MoH servers, these are stored as structured data in the local EMRs.

In the NHIS-T, two standards are important: (1) the United Nations Centre for Trade Facilitation and Electronic Business (UN/CEFACT) Core Component Technical Specification (CCTS) to design EHR content in the conceptual base, and (2) HL7 CDA to serialize this conceptual design.

Design of the Electronic Health Record Content

UN/CEFACT CCTS is a methodology to define the structured, abstract document components used to increase the interoperability of electronic business documents. In the NHIS-T, EHR content is designed based on UN/CEFACT CCTS to assure the reuse of common information blocks in EHRs. First, the data types and the data elements used in EHRs are

identified, and then a set of reusable building blocks of the EHRs, named the Minimum Health Datasets (MHDS), is produced. Some examples of the MHDS are maternal mortality dataset, diabetes dataset, dialysis patient dataset, patient admission dataset, cancer dataset, chronic disease dataset, etc.

In every dataset, there are many data elements. For example, data elements of cancer datasets are data of first diagnosis; diagnostic method; location histological type; the Surveillance, Epidemiology, and End Results Program (SEER) summary stage; laterality; occupation; and cancer.

The data elements are coded with universal medical terminologies (eg, International Classification of Diseases and Health Related Problems version 10, ICD-10; Anatomical Therapeutic Chemical, ATC; etc) and predefined categorical values standardized by the MoH, such as gender or marital status. All kinds of these terminologies are selected by the MoH and available from the Health Coding Reference Server (HCRS). There are 342 code systems in the HCRS; the current version of the HCRS is 3.0 and is available on the Internet via Web services. A tabular representation is also available on an official Web page and allows users to query by means of Web browsers.

These reusable building blocks (MHDS) are assembled into aggregate document components called Transmission Datasets (TDS, episodic EHRs). Some of the TDS are physical examination TDS, laboratory test results TDS, and in-patient TDS.

All the data elements, MHDS, and TDS are identified by the MoH in the light of the needs of stakeholders (eg, strategic decision makers, health care organizations, academic institutions, etc) and published in the National Health Data Dictionary (NHDD). This is a dynamic and continuously improving process. When required, new MHDS are produced using existing data elements, and the NHDD is improved by identifying new data elements. The total number of data elements, MHDS, and TDS in the most recent version of NHDD; namely, version 2.2, are 418, 66, and 7, respectively. All versions of NHDD are available on the official website of The Turkey e-Health project [14].

Transport of the Electronic Health Record Content

This conceptual EHR architecture is serialized into extensible markup language (XML) based on the HL7 CDA structure. As described in the first part of this miniseries, HL7 CDA is a document mark-up standard referred to in the exchange of information as part of the HL7 version 3 (V3) standards that aim to specify the structural and semantic aspects of clinical documents [15].

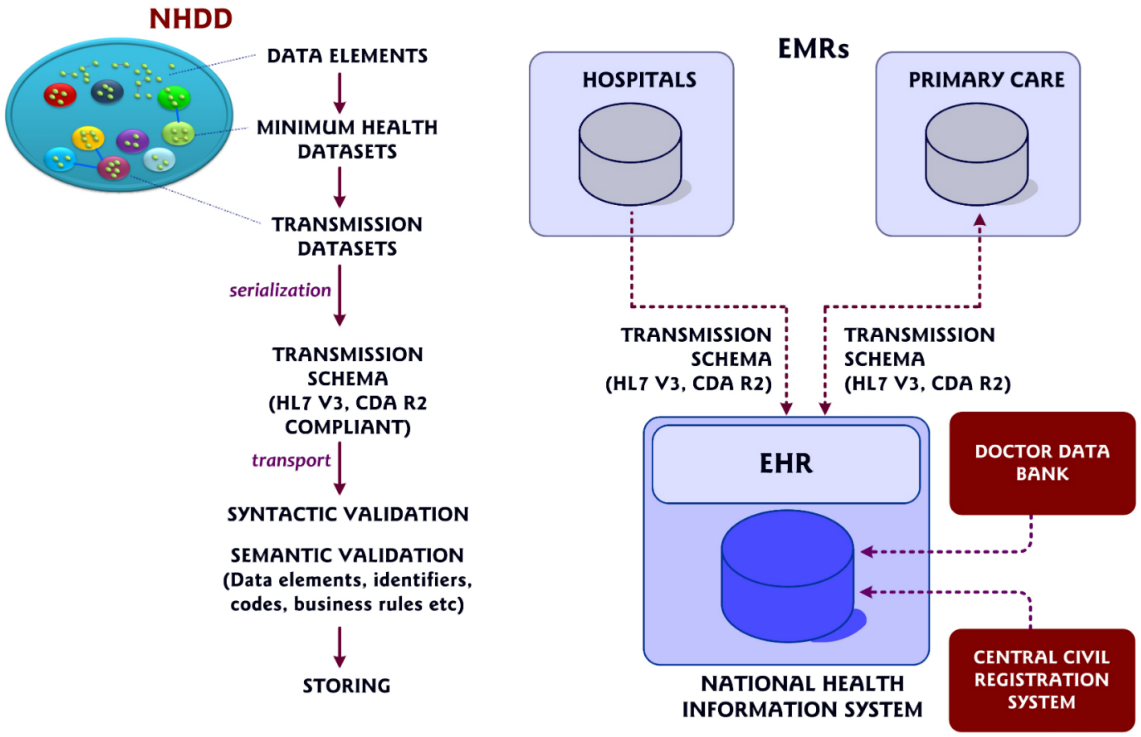
In the serialization process, the TDS are mapped to HL7 CDA to create the “transmission schema”. Each transmission schema is wrapped with a root element named after the main dataset in the transmission [12,13].

The NHIS-T messaging system accepts HL7 V3 CDA release 2 (R2) as a reference and is compliant to this standard. Therefore, the messages sent must comply with the rules defined in the CDA XML Schema Definition file. Health care organizations send messages containing clinical data to the central MoH servers through Web services. In the current version of the NHIS-T system, the transmission schema instances are localized according to Turkey’s HL7 Profile. During this process, the rules, which are set in the “HL7 Refinement, Constraint, and Localization” document, are applied. All of these templates are created and published by the MoH, and used as the standard by health care information systems vendors [12,16].

Incoming messages are validated regarding syntax and semantics, and the appropriate messages are stored in the NHIS-T central repositories. Patient and medical professional identifiers are acquired and validated from the Central Civil Registration System and Doctor Data Bank, respectively (Figure 1 shows the NHIS-T) [13,16].

The current version of the NHIS-T messaging system allows for the transfer of medical data from health care providers’ (hospitals and family practitioners) information systems to central servers via Web services. It has the infrastructure that will provide access to patients’ records for authorized health care professionals within the hospital, which will allow patients to reach their own medical data, for example, personal health records (PHRs). But, the legal regulations have to be completed before both types of access—authorized or self—are available. Then, the establishment of a PHR system will be allowed [13].

Figure 1. Schematic representation of National Health Information System of Turkey. NHDD=National Health Data Dictionary; HL7 CDA R2=Health Level 7 Clinical Document Architecture release 2; EMR=electronic medical record; and EHR=electronic health record.



Architectural Extensions for Single Nucleotide Polymorphism Data Incorporated National Health Information System of Turkey

It is necessary to build an infrastructure providing clinicogenomic information and subsequent updates to the physicians. A curated knowledge base extracting clinical information from relevant SNP data, and supporting systems processing up-to-date data for clinical decisions over the patient's lifetime should be integrated as clinicogenomic information into medical records [17,18].

In the light of our literature review, for a genome-enabled NHIS-T, improvements in three components are needed: (1) enhancement of existing messaging infrastructure to share personal SNP data and clinicogenomic associations between stakeholders, (2) development of a national-level ClinGenKB for transforming personal SNP data to clinicogenomic associations, and (3) advancement of end-user applications (EMR, PHR, etc) for reporting of clinicogenomic interpretation of clinically relevant SNP data.

A messaging infrastructure ensures the connection of different stakeholders and the sharing of all types of relevant data among these partners, for example, relevant SNP data between genomic laboratory and knowledge base, and clinicogenomic associations between knowledge base and end-user applications.

As clinicians cannot extract the clinical interpretation of all SNP variations directly from the medical sources due to temporal and cognitive limitations, the integration of the clinical interpretations of variations (eg, clinicogenomic associations) into the medical record will be more efficient for clinical decision making [19-22]. To convert SNP data into clinicogenomic associations and infer clinically meaningful results, we need a structured knowledge source, for example, a knowledge base.

Most of the clinically relevant SNPs have minor effects (odds ratio <1.50-2.00), and there are only limited numbers of different examples. For example, in our study, we extracted more than 209 prostate cancer-associated SNPs from the literature, and many of these SNPs have minor effects for predicting the prostate-cancer risk [9]. Additionally, in many cases, SNPs do not show their effects directly for a given disease, but do so in combination with other SNP variations and clinical and environmental factors, which require a much higher level of probability calculations.

Previously, various approaches were proposed to assess and report clinicogenomic associations. The listing of clinicogenomic associations and their effects may be useful for a limited number of independent associations. But, it is clear that clinicians cannot interpret and evaluate all variations individually, especially for SNPs with small impact degrees.

Each day, the volume of variation data integrated into clinical practice exceeds the boundaries of unsupported human cognition and interpretive capacity. Additionally, the rapidly growing literature on clinicogenomic associations increases the challenge for professionals to stay current.

Therefore, the polygenic risk models that are under development or panels, which assign values for various SNP alleles and calculate the total risk of diseases, will be more effective for risk prediction. Eventually, we also need end-user applications that report clinicogenomic associations independently and risk-value calculations based on predictive models.

Regarding technical capabilities (eg, network bandwidth, storing and processing capacities, etc), different types of architectures can be developed, but the development of two additional components (knowledge base and reporting capability) is essential. A ClinGenKB must be constructed at the national level as a manually curated and continuously updated source that would include clinical information and its possible associations with SNP variants. In the end-user, decision-support applications (EMRs, PHRs, etc), clinicogenomic associations, and external data (eg, family history) must be interpreted independently, or based on predictive models to support decision making.

In the existing NHIS-T, medical and laboratory test results are sent from hospitals to the central EHR databases as "Examination Result Transmission Dataset". The HL7 CDA R2 conformant transmission schema of this dataset includes several MHDS, for example, registration MHDS, result of tests MHDS, patient MHDS, etc. "Result of tests MHDS" involves data elements about examination features (order time, protocol number, result time, test result, reference value ranges, etc). The data type of laboratory analysis should be numeric or textual data regarding current schema standards. HL7 V3 interoperability standards support encapsulated data type for text data [15].

Although whole genome sequencing (WGS), whole exome sequencing (WES), and other types of genotyping tests are accepted as laboratory tests, they have different characteristics than other laboratory tests in routine practice. After a clinical WGS/WES test, a personal SNP data file which contains a large amount of variant data is produced, in which all variant data needs to be managed in an effective way. In the proposed architecture, personal sequencing data is acquired and stored as raw data within a genomic laboratory information system. The clinically relevant personal SNP (CR-SNP) data is extracted from a personal SNP data file using the CR-SNP data list (genomic identifiers of clinicogenomic associations in the ClinGenKB). Then, a personal CR-SNP data file is sent via the NHIS-T infrastructure from a genomic laboratory to central EHR databases as an encapsulated text file. This file has to include SNP identifiers, for example, reference SNP identifier (rsID) and allele data in the HL7 CDA R2 schema (Figure 2 shows this process).

The received CR-SNP files would be stored within the central EHR databases. Then, the CR-SNP files can be processed to infer clinically relevant data by using the clinicogenomic associations from the knowledge base. Resulting personal, clinicogenomic association files would be sent to end users. As shown in Figure 2, based on existing technical capabilities, to decrease the load of sharing clinicogenomic association files, a replicated knowledge base could be integrated as a Web application running on the client side, whereas a CR-SNP data file could only be stored in central servers. In this situation, clinicogenomic associations are inferred at the client side through the replicated knowledge base. In that case, the client-side knowledge base must be frequently synchronized with the central knowledge base.

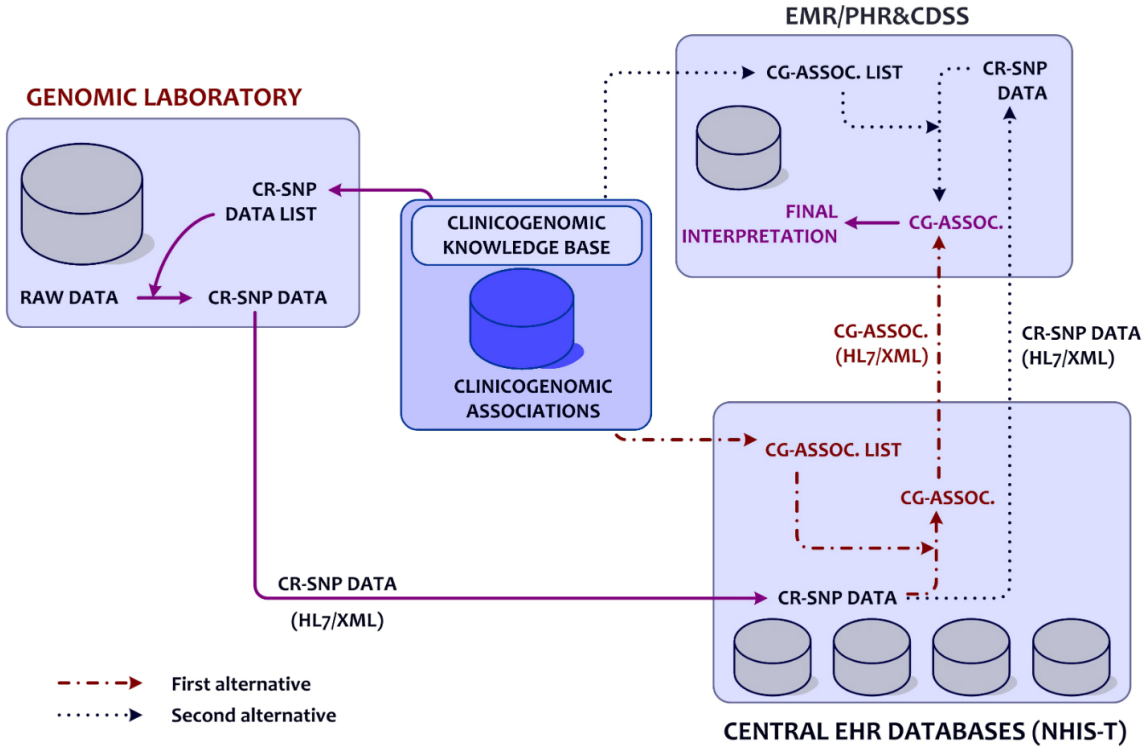
To link personal CR-SNP data and clinicogenomic associations, the rsID and allele combination is used. Data input to the knowledge base rules (or associations) require rsID and allele information, and medical interpretation, significance, and representative information are sent back as the output from the knowledge base. The rule structure is explained in the section on "Definition of Clinicogenomic Associations". Finally, personal clinicogenomic associations are inferred from personal CR-SNP data and a knowledge base.

When an authorized user (patient, family practitioner, or a medical specialist) needs to reach a personal CR-SNP or a clinicogenomic association file, a request should be sent to the central EHR, and a current data file would be received via the NHIS-T communication infrastructure.

Additionally, it is recommended that an independent medical authority, established by domain experts, should update ClinGenKB. According to the type and level of change and preferred architecture, existing personal SNP data, CR-SNP data, and/or clinicogenomic associations must be reinterpreted after the authorization of the patient, through the genomic laboratory system at the national EHR repository and/or the client side.

Our assessment of the NHIS-T reveals that its capabilities (eg, regarding Web services, client-side inference, and reporting capabilities, PHRs, if requested) need to be extended to be able to share CR-SNP or personal clinicogenomic associations between central EHR databases and end-user systems. Here, we are presenting complementary capabilities developed as prototypes for the NHIS-T, for example, ClinGenKB and ClinGenWeb, which specifically focus on disease risk assessment. In our study, as an initial attempt through the development of much sophisticated infrastructure, we have concentrated on the interpretation of SNP variant data and excluded other types of variants. The use of personal clinicogenomic information to determine the disease risk of a patient's family members is considered to be out of scope. Also, security and privacy issues, as well as constraints about hardware and infrastructure, are excluded.

Figure 2. Alternatives for extended architecture of genome enabled National Health Information System of Turkey (NHIS-T). CR-SNP=clinically relevant single nucleotide polymorphism; HL7=Health Level 7; XML=extensible markup language; CG-ASSOC.=clinicogenomic associations; EMR=electronic medical record; PHR=personal health record; and CDSS=clinical decision support system.



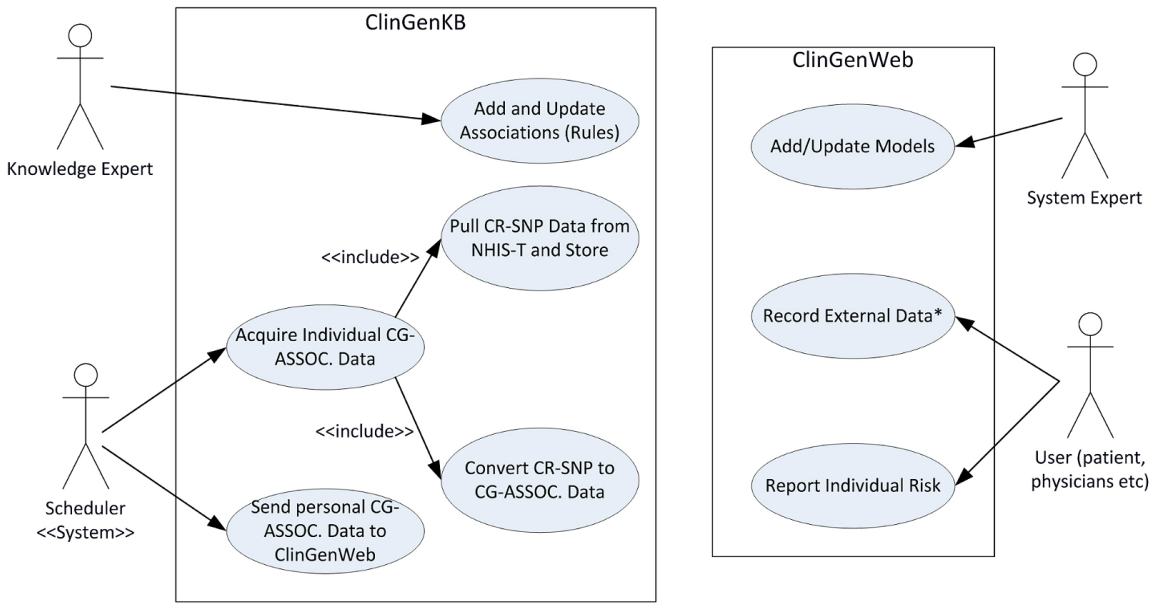
Design of the Complementary Components

General Use Case

Figure 3 shows the general use case diagram of our approach. Major actors of our system are the end-users (physicians or patients), knowledge expert, and system expert. The NHIS-T infrastructure, ClinGenKB, and ClinGenWeb will perform the sending and storing functionalities. Knowledge experts add and update clinicogenomic associations into the ClinGenKB. The conversion process of the CR-SNP to the clinicogenomic associations is accomplished synchronously by ClinGenKB at the first uploading and then—as a rule—in update sessions.

Before the design and development of the ClinGenKB and the ClinGenWeb, we standardized the clinicogenomic associations and models, as explained below.

Figure 3. The use case diagram of our Clinicogenomic Knowledge Base (ClinGenKB) and Clinicogenomic Web Application (ClinGenWeb). CR-SNP=clinically relevant single nucleotide polymorphism; CG-ASSOC.=clinicogenomic associations; NHIS-T=National Health Information System of Turkey; and *External Data: environmental, behavioral, family health history data, etc.



Standardization of the Clinicogenomic Approaches

At the end-user side, assessment and reporting of clinicogenomic associations using several approaches, for example, listing of the independent associations, complete visualization of independent associations, polygenic risk scoring, and model-based methods, etc, were determined as requirements of the system. The common data fields and approach-specific additional data fields of clinicogenomic associations in the ClinGenKB must be defined for handling a variety of approaches. In particular, the ClinGenKB should collect all types of independent and model-based clinicogenomic associations as a whole set. So, in our study, we reduced the independent and different types of model-based associations into a standard definition while designing the ClinGenKB.

In the scientific literature, we have determined two types of models, for example, cumulative and probabilistic models, for prostate cancer risk assessment. In cumulative models, we can calculate the possible disease risk by combining the impact of several clinicogenomic associations. In a probabilistic model, SNP profiles with increased disease risk are determined with an evidence-based approach during development of the model, and the patient's risk is determined through the patient's genotyping profiles. Detailed examples of these models will be explained in the next section of the miniseries [9]. Examples of the cumulative model of prostate cancer and their reference tables are given in Multimedia Appendices 1 and 2, and examples of the probabilistic model with the list of the possible SNP profiles are provided in Multimedia Appendices 3 and 4.

Through analysis of these models, we have extracted possible clinicogenomic association types for assessment and reporting. Finally, we have generated a comparative table to determine the requirements for the design of the ClinGenKB (Table 1).

Table 1. Comparison of representation types and definitive data fields.

Definitions		Independent associations	Complete visualization for independent associations	Associations of cumulative models	Associations of probabilistic models
Association					
	rsID	X	X	X	X
	Allele	X	X	X	X
	Disease code	X	X	X	X
	Disease name	X	X	X	X
	Magnitude of impact	X			
	Degree of evidence quality	X			
	Impact category		X		
	Evidence category		X		
	Impact value				
	Branch_id[c]				X
Model					
	Model type	X	X	X	X
	Model name	X[a]	X[a]	X	X
	Total impact			X	
	Total count of SNPs				X
	Branch_id				X
	Narrative interpretation			X	X
External data					
	Family history			X	X
	Other type[b]				X

[a]In this study, only increased risk covered

[b]BMI = body mass index, alcohol consumption, smoking, etc

[c]Branch_id, a numeric identifier for every association which is derived from a probabilistic model.

Definition of the Clinicogenomic Association

Next, we have produced a standard representation for all types of clinicogenomic associations, including an association identifier, genomic parts, and clinical parts. Genomic parts contain SNP data, namely, rsID and allele values. In the clinical part, there are medical data, model information, values about impact, and evidence degree of per association.

Detailed data analysis of association typology is presented in Table 2. This table was used while designing the ClinGenKB.

Table 2. Data field analysis of association parameters.

Data category	Parameters	Values
Association identifier	Assoc_id	Unique numeric value
SNP data	rsID	rs value
	Allele	Allele value (eg, AA[e]; AT[f]; , CG[g]; etc)
Medical data	Disease code and name	ICD-10
Representation model of associations	Model type	Independent associations, cumulative model, and probabilistic model
	Model name	Increased[a] ; (number of SNP and name of first author)[b]; (model number, name of authors, and date)[c]
Association values	Parameter 1 (magnitude of impact[a]; impact value[b]; and branch_id[c])	Numeric value[a, b]; numeric identifier[c]
	Parameter 2 (degree of quality of evidence[a])	Numeric value (between 1 and 3)
	Parameter 3 (impact category[d])	1,2,3 (corresponding Weak, Moderate, Strong, respectively)[d]
	Parameter 4 (evidence category[d])	1,2,3 (corresponding Weak, Moderate, Strong, respectively)[d]

[a]Independent associations

[b]Cumulative model-based associations

[c]Probabilistic model-based associations

[d]Complete assessment for independent associations

[e]AA = adenine-adenine

[f]AT = adenine-timine

[g]CG = cytosine-guanine

Data Fields of a Clinicogenomic Association

In this representation, clinicogenomic associations must have a unique identifier assigned automatically. In ClinGenKB, SNPs are identified by both rsID and allele data that correspond to the forward strand of genomic sequence. If a SNP is related with the different medical conditions or models, for every instance, a new association was defined, and a different unique identifier was assigned.

The medical data category contains diagnosis codes and names. Values from this data field are selected from the ICD-10, which is used for diagnostic terminology for diseases in the current NHIS-T.

Model data has two components: (1) type, and (2) name. In ClinGenKB, we have two main clinicogenomic associations, for example, model-based or model-free (or independent) associations. Names for independent associations are categorized as increased or decreased regarding the potential risks and/or protective characteristics. In this study, we have focused only on increased risk. Model-based associations are used in the predictive models.

Association values are tightly related to the type of model, for example, independent associations, cumulative model, and probabilistic model-based associations. For an independent associations odds ratio (OR), the degree of evidence quality, impact values, and evidence categories were found to be appropriate and sufficient elements to evaluate clinical significance, both individually and as a whole. In cumulative model-based associations, it is necessary to assign an impact value for every association to calculate the total personal risk value according to the model-definition table. For the probabilistic model, we have calculated the total effects of variants using their branch_id. On the client side, all possible associations derived from the probabilistic model are grouped by “branch_id”, and for all of these groups, the total impact of risk parameters is determined. If one of these values is equal to the total value of the corresponding branch, it is interpreted as the patient having the risk of prostate cancer, based on the accuracy, precision, and recall values of the model. For example, in the only SNP model, 154 different possibilities were defined. According to the first branch (branch_id=1), if an individual carries all of the “rs11720239-AA, rs2999081-CT, rs2811518-CT, and rs4793790-TT” SNP variations, it is assessed as this individual having a risk of developing prostate cancer

with the degree of the accuracy, precision, and recall of the model.

Predictive Risk Models

Additionally, a standardized model definition table involving reference values for variants and corresponding disease risks is produced, as the final interpretation of clinicogenomic associations will be completed at the end-user side (ClinGenWeb) using predictive models.

We have developed the model definition table for the analysis of the model-based associations on the client side (Table 3). The model type identifies the category of models, and model name labels them. The total value and explanation fields are mapped to the total impact of related SNPs and the corresponding risk categories. For cumulative models, these fields are about total impact and its explanation. For the probabilistic models, total value is referred to as the count of all SNPs for every branch; the explanation is the interpretation of risk values regarding accuracy and precision. An additional data field identifying the branch_id of the selected support vector machine- iterative dichotomiser 3 hybrid model is needed for the explanation of the probabilistic model.

Examples of cumulative models are given in Multimedia Appendices 1 and 2, and a list of examples of probabilistic models and their parameters is given in Multimedia Appendices 3 and 4.

Table 3. Data field analysis of model definition table. OR: odds ratio.

Parameters	Value (domain)	Explanation
Model type	Cumulative model, probabilistic model	
Model name	(Number of SNP and name of first author)[a]; (model number, name of author, and date)[b]	
Total value	Numeric value	Total impact[a]; total count of SNPs[b]
Explanation 1	Text value	Explanation (OR)[a]; Explanation (brief interpretation about risk assessment)[b]
Explanation 2	Text value	Explanation (branch_id)[b]

[a]Cumulative model-based associations

[b]Probabilistic model-based associations

Development of the Complementary Components

Clinicogenomic Knowledge Base

Knowledge bases are repositories that help to collect, organize, share, search, and utilize information. Developing an accurate, accessible, structured clinicogenomic knowledge source (ClinGenKB) for disease-associated SNPs (prostate cancer in our case) is an essential component of the proposed clinicogenomic information-integrated EHR.

Raw genomic variant data is not appropriate to support clinician decision-making due to its high dimension. The clinical association of the variant is convenient to transfer for clinical decision support, where the interpretation of the variant and its associated clinical meaning is periodically updated in the knowledge base. The data field tables described in Figure 3 must be kept up to date and shared with other stakeholders to be used as a standard reference for the interpretation of the model-based associations in a proper manner. Also, model type and model name fields must be used as standard references for the same fields in the definition tables. Such a system would allow for the reinterpretation of variant data throughout dynamic updates.

Technically, there are many tools for knowledge modeling and implementation. For this study, we have preferred to develop our prototype using BioXM, which is a distributed software platform providing a central inventory of information and knowledge [10]. Through BioXM, we were able to quickly generate, easily manage, and visualize the scientific models as extendible networks of interrelated concepts.

To develop the ClinGenKB on the BioXM platform, we have designed the domain-specific data model with semantic objects—elements, annotations, ontologies, and databanks—and the connections (relations) using the BioXM graph viewer based on our clinicogenomic association definitions. Next, we have defined the importing scripts to transfer the extracted independent and model-based clinicogenomic associations and the personal CR-SNP data to the knowledge base. BioXM supports the data import and export as XML, hyper text markup language, Excel, or plain text format. Finally, we have prepared views, queries, and smart folders to manage our data model and the inferring processes.

Our domain model defining elements, annotations, relations, and scope of these components in BioXM is based on the association definition table (Figure 4 shows this table). In this domain model, we have three types of elements: (1) person, (2) SNP variant, and (3) clinical association. Every element has its specific annotations. The personal element is related with the SNP variant by a “has” relation, referring to the fact that each patient will have a set of SNP variants. SNP identifiers are assigned to variants for ensuring uniqueness. Then, each SNP variant is related with a clinical association element as the input.

We have imported the content of the knowledge-based definition table as an external file with scripts. This content can be updated with subsequent importing operations. If a new association is generated or existing associations are changed or cancelled,

authorities can organize all the changes in an external source according to the association definition table, and then can easily upload all of them via BioXM compatible files. After the importing process, the clinicogenomic associations can be sorted and managed by system administrator from table (Figure 5 shows a screenshot).

In addition, we can store personal CR-SNP data as a separate file on BioXM. CR-SNP data can be easily converted to clinicogenomic associations based on the content of ClinGenKB, and this data file is exported as a text file. For all individuals whose CR-SNP data is stored in BioXM, whenever it is needed, it is possible to access personal CR-SNP data, and to produce new clinicogenomic association data files based on the current ClinGenKB. In the NHIS-T, all of these personal files can be accessible with the inclusion of the Turkish citizen-identifier number in our prototype, and can be sorted according to data categories (Figure 5).

Figure 4. The graphical representation of designed prototype with BioXM Knowledge Management Environment. URL=uniform resource locator; rsID=reference single nucleotide polymorphism identifier; and SNP=single nucleotide polymorphism.

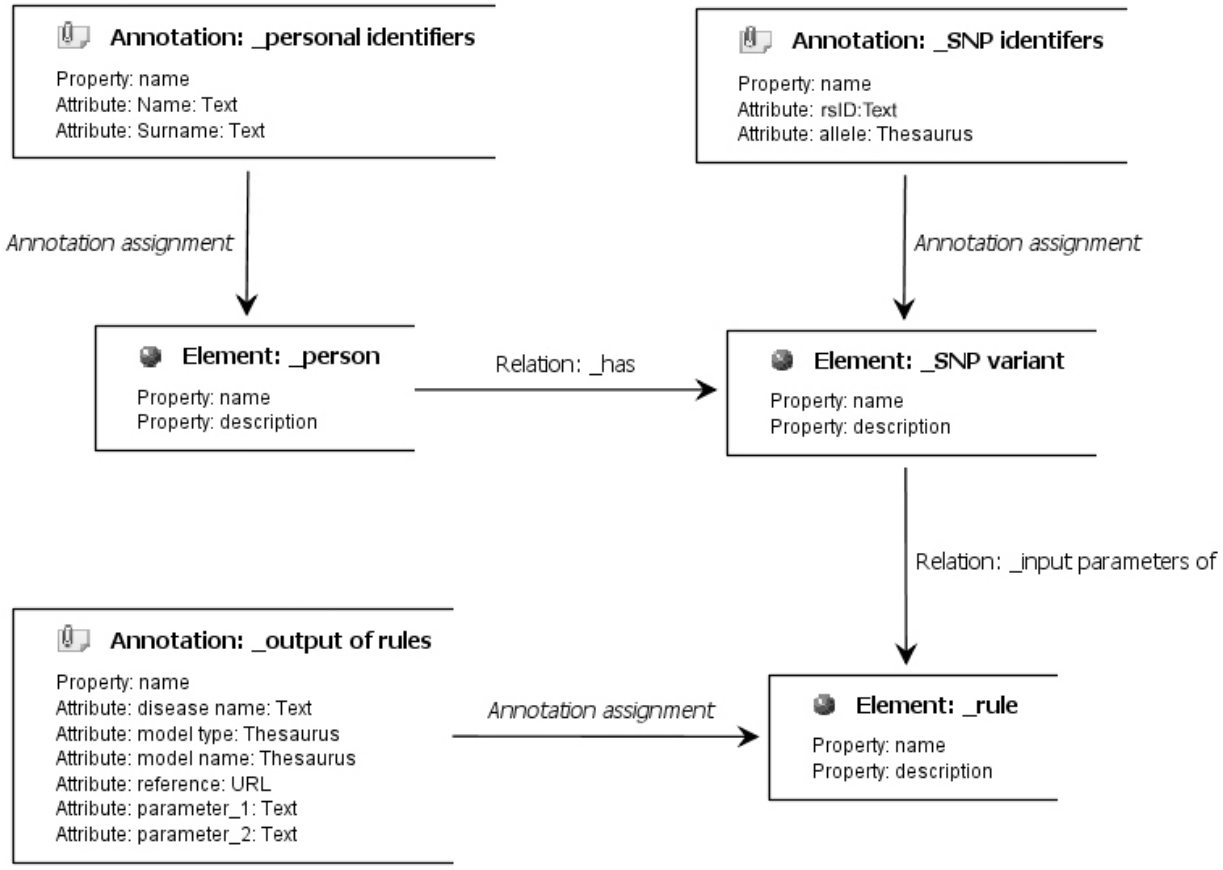
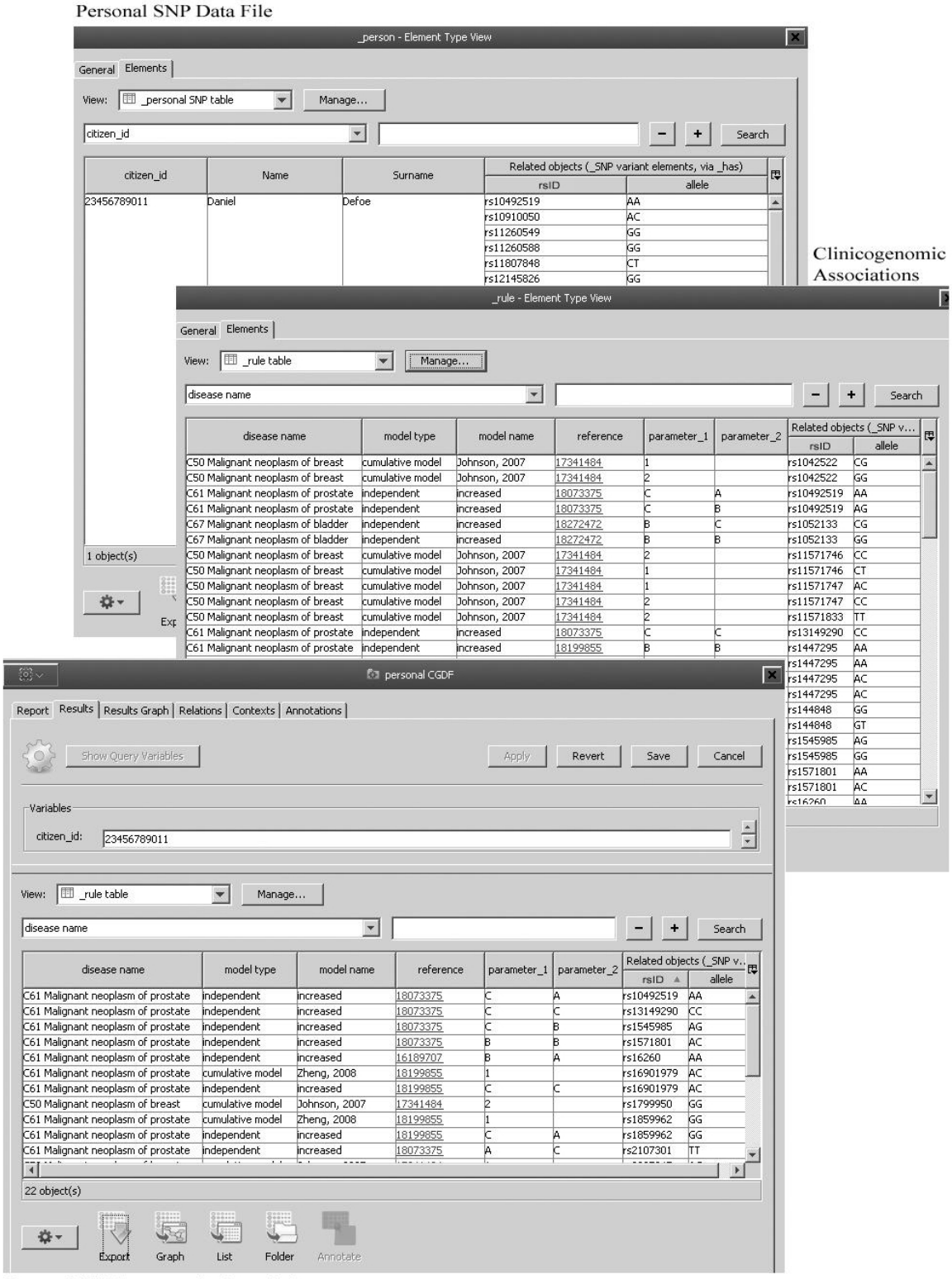


Figure 5. Functional steps for Clinicogenomic Knowledge Base (ClinGenKB): (1) defining clinicogenomic content, (2) uploading personal clinically relevant single nucleotide polymorphism (SNP) data, and (3) inferring personal clinicogenomic associations based on the content. rsID=reference single nucleotide polymorphism identifier; AA=adenine-adenine; AC=adenine-cytosine; AG=adenine-guanine; CC=cytosine-cytosine; CG=cytosine-guanine; CT=cytosine-thymine; and GG=guanine-guanine.



Clinicogenomic Web Application

After the transmission of the personal, clinicogenomic association data file to the end-users' (medical specialists, family practitioners, and patients) applications, another critical issue is the final interpretation and reporting of the results. Reporting presents itself here as a critical point for maximizing the effectiveness of the overall system in translating clinicogenomic data into the clinic. High-dimensional variant data and its clinical associations—along with its interpretations—have to

be reported and visualized in a simplistic and holistic manner for easy interpretation by both health care professionals and patients.

Regarding clinicogenomic decision support, our approach aims to divide clinicogenomic interpretation into two phases, namely: (1) conversion of the variant SNP into a clinicogenomic association, and (2) clinical interpretation of these associations. Final interpretation is completed on the client side. This approach gives us the flexibility to add/update external parameters that will be monitored or collected by end users. For example, in some cumulative prostate models, positive family history augments the total risk value in addition to clinically relevant SNPs. Family history is a dynamic parameter that can change in time. Patients ideally accomplish effective tracking of changes in family history. Similarly, clinical, environmental, behavioral, or sociodemographic factors, should be involved to assess the total risk with variant data at the end-user level.

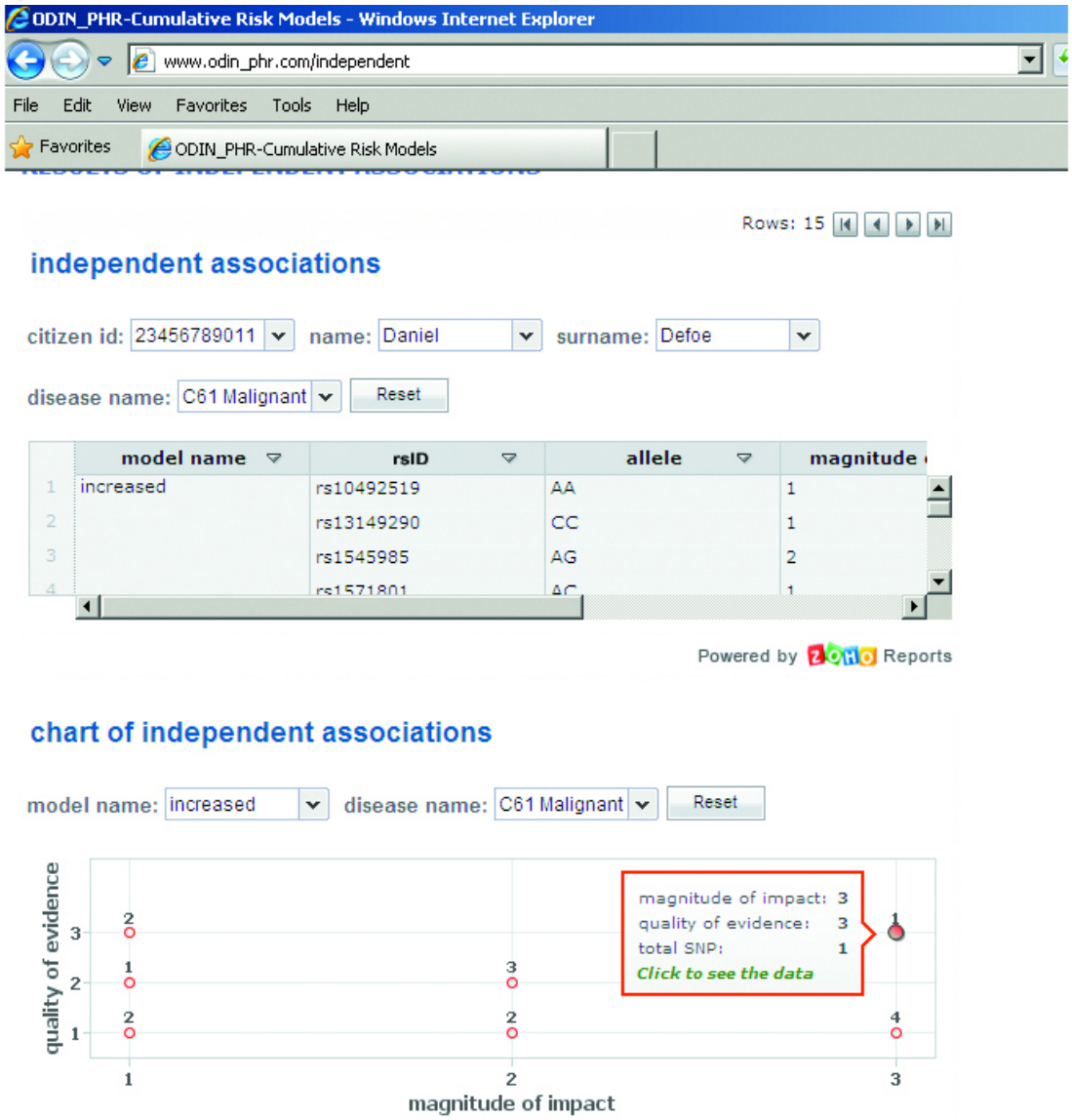
Accordingly, we have developed a practical reporting approach and demonstrated it using Zoho Reports as a prototype system (namely, ClinGenWeb) for the client side. Zoho Reports is an on-demand reporting and business intelligence tool that supports several, report generation capabilities, for example, chart/graph, tabular views, summary views, pivot tables, dashboards, and structured query language (SQL)-driven querying. Most importantly, it is possible to embed generated reports within external Web sites and Web applications [11]. ClinGenWeb is developed as a Web application, processing genomic associations and clinical and environmental risk parameters. ClinGenWeb is designed with the capability to report relevant clinicogenomic SNPs or to assess individual risk based on different models with the combination of conventional health data and clinicogenomic associations.

The ClinGenWeb is designed to report personal predictive risk under three main categories: (1) detailed reporting of individual associations, (2) the assessment of a number of clinically relevant SNPs, and (3) model-based interpretations of the clinicogenomic associations. Basic models are based only on assessing the relevant SNPs, whereas other predictive models require additional clinical data (family history, BMI, etc). When provided, corresponding risk factors for prostate cancer can be used to calculate the model-based risk. Also, external personal data about clinical and some environmental risk factors for prostate cancer can be reported.

Reporting of Independent Associations

The reporting of all independent associations individually would be very confusing, and the interpretation of this data by users would be time consuming. So, the associated data are presented in a category-based graph, where the x- and y-axes correspond to impact and evidence categories, respectively (Figure 6 shows this graph).

Figure 6. Visualization of independent associations in the Clinicogenomic Web Application (ClinGenWeb). Independent associations and their clinical significances are listed as a table and represented on a category based graph. rsID=reference single nucleotide polymorphism identifier; SNP=single nucleotide polymorphism; AA=adenine-adenine; CC=cytosine-cytosine; and AG=adenine-guanine.

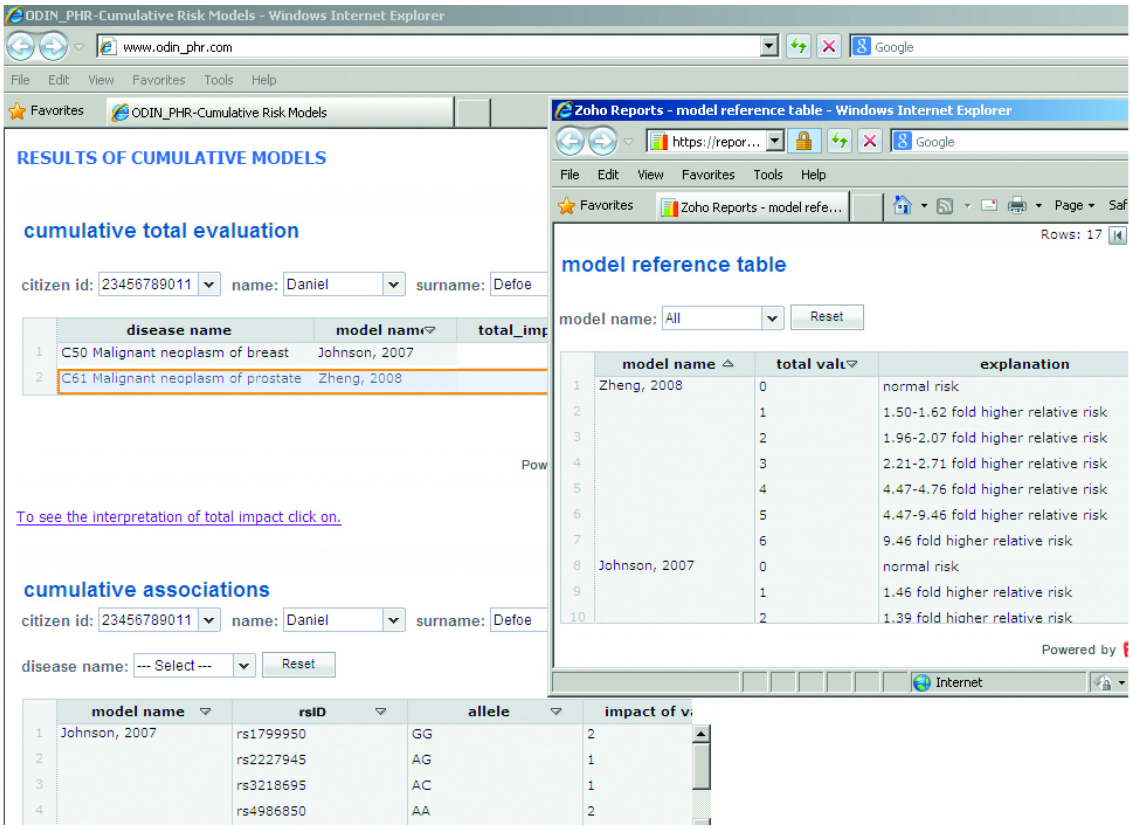


Reporting of Model-Based Associations

Comparatively, a model-based interpretation provides us with more effective information for supporting the end-users' decision making. This type of rule is based on accepted and proven integrated models as described in the next section.

In our study, we have used two kinds of models, namely, cumulative and probabilistic models. In the ClinGenWeb, the results of these models and detailed explanations of reference values are presented to the end-users as a complete set of information. If needed, end-users can exploit the detailed analysis of risk factors as in the total evaluation of the model (Figure 7 shows this analysis).

Figure 7. Reporting and interpretation of the results of model-based associations and the whole impact (and meaning) of models in the Clinicogenomic Web Application (ClinGenWeb). rsID=reference single nucleotide polymorphism identifier; AA=adenine-adenine; AC=adenine-cytosine; AG=adenine-guanine; and GG=guanine-guanine.



Combining Clinicogenomic Associations With the External Data

In ClinGenWeb, users can record and store additional types of risk factors (family history, environmental, behavioral, and clinical data) to assess the increase in their prostate cancer risk. When additional data is collected, it can be used as a parameter for the model or just included in the final report.

Discussion

In this article of the miniseries, we have presented the design and development of the required structures for the NHIS-T to incorporate SNP and clinicogenomic data for disease risk assessment in the light of the first part of the miniseries, and of the analysis of the existing NHIS-T.

We have proposed the possible architectures and extensions of HL7 CDA templates to transmit personal, clinically relevant SNP data and clinicogenomic associations. In this approach, two important complementary capabilities are the structured knowledge base and the end-user assessment and reporting applications. The knowledge base (ClinGenKB) is responsible for transforming personal, clinically relevant SNP data to meaningful clinicogenomic associations. The end-user application (ClinGenWeb) ensures the issuing of personal reports where extracted clinicogenomic associations are listed, visualization of the risky SNPs, and the calculation of the total risk based on proposed risk models.

In the next part of this article, we will focus on the extraction of SNP associations to build the proposed ClinGenKB, and on the evaluation of the proposed components for the NHIS-T for determining prostate cancer risk using real direct-to-consumer SNP data files. In addition, assessment and reporting approaches to calculate personal prostate cancer risk will be presented.

Acknowledgments

We would like to thank Biomax Informatics AG for their permit, help, and support to use BioXM Knowledge Management Environment.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Cumulative Models for Prostate Cancer.

[XLSX File (Microsoft Excel File), 11KB - medinform_v2i2e17_app1.xlsx]

Multimedia Appendix 2

Reference Tables for Cumulative Models.

[XLSX File (Microsoft Excel File), 10KB - medinform_v2i2e17_app2.xlsx]

Multimedia Appendix 3

List of First Probabilistic Model Based Associations (Only SNP Model).

[XLSX File (Microsoft Excel File), 27KB - medinform_v2i2e17_app3.xlsx]

Multimedia Appendix 4

List of Second Probabilistic Model Based Associations (SNP-Environmental Combined).

[XLSX File (Microsoft Excel File), 11KB - medinform_v2i2e17_app4.xlsx]

References

1. Ginsburg GS, Willard HF. Genomic and personalized medicine: Foundations and applications. Transl Res 2009 Dec;154(6):277-287. [doi: 10.1016/j.trsl.2009.09.005] [Medline: 19931193]
2. Chan IS, Ginsburg GS. Personalized medicine: Progress and promise. Annu Rev Genomics Hum Genet 2011;12:217-244. [doi: 10.1146/annurev-genom-082410-101446] [Medline: 21721939]
3. Scheuner MT, de Vries H, Kim B, Meili RC, Olmstead SH, Teleki S. Are electronic health records ready for genomic medicine? Genet Med 2009 Jul;11(7):510-517. [doi: 10.1097/GIM.0b013e3181a53331] [Medline: 19478682]
4. Belmont J, McGuire AL. The futility of genomic counseling: Essential role of electronic health records. Genome Med 2009;1(5):48 [FREE Full text] [doi: 10.1186/gm48] [Medline: 19439060]
5. Garets D, Davis M. HIMSS analytics. 2006. Electronic medical records vs electronic health records: Yes, there is a difference URL: http://www.himssanalytics.org/docs/WP_EMR_EHR.pdf [accessed 2013-12-03] [WebCite Cache ID 6Lb0ZAXxG]
6. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. Int J Med Inform 2008 May;77(5):291-304. [doi: 10.1016/j.ijmedinf.2007.09.001] [Medline: 17951106]
7. Healthcare Information and Management Systems Society (HIMSS) Global Enterprise Task Force. HIMSS. Electronic health records: A global perspective, part 1 URL: http://www.himss.org/files/HIMSSorg/content/files/Globalpt1-edited%20final.pdf [accessed 2013-12-03] [WebCite Cache ID 6Lb0VPAh0]
8. Beyan T, Aydın Son Y. Incorporation of personal single nucleotide polymorphism (SNP) data into a national level electronic health record for disease risk assessment, part 1: An overview of requirements. JMIR Med Inform 2014 Jul 24;2(2):e15. [doi: 10.2196/medinform.3169]
9. Beyan T, Aydın Son Y. Incorporation of personal single nucleotide polymorphism data into a national level electronic health record for disease risk assessment, part 3: An evaluation of SNP incorporated NHIS-T for prostate cancer. JMIR Med Inform (forthcoming) 2014 (forthcoming). [doi: 10.2196/medinform.3560]
10. BioXM knowledge management environment. Germany: Biomax informatics AG, the life sciences knowledge management company URL: http://www.biomax.com/products/bioxm-knowledge-management-environment/ [accessed 2014-07-29] [WebCite Cache ID 6RQN6Qfyc]
11. Zoho Corporation. Zoho reports, online reporting and business intelligence. 2013. 2013 URL: https://www.zoho.com/reports/ [accessed 2014-07-29] [WebCite Cache ID 6RQNHDOap]
12. Kabak Y, Dogac A, Kose I, Akpınar N, Gurel M, Arslan Y, et al. The use of HL7 CDA in the national health information System (NHIS) of Turkey. In: Proceedings of 9th International HL7 Interoperability Conference. 2008 Oct Presented at: 9th International HL7 Interoperability Conference; 8-11 October 2008; Crete, Greece URL: http://www.srdc.com.tr/publications/2008/FinalCDA.Turkey.IHIC08Formatted.pdf
13. Dogac A, Yuksel M, Avci A, Ceyhan B, Hülür U, Eryilmaz Z, et al. Electronic health record interoperability as realized in the Turkish health information system. Methods Inf Med 2011;50(2):140-149. [doi: 10.3414/ME10-01-0022] [Medline: 21132219]
14. Republic of Turkey Ministry of Health. National health data dictionary 2. (In Turkish) URL: http://www.e-saglik.gov.tr/dosyalar/USVS2_30032012.pdf [accessed 2013-12-03] [WebCite Cache ID 6Lb0x4RML]
15. Benson T. Principles of health interoperability HL7 and SNOMED (Second Edition). London: Springer-Verlag; 2012.

16. Kose I, Akpinar N, Gurel M, Arslan Y, Ozer H, Yurt N, et al. Turkey's national health information system (NHIS). In: Proceeding of the eChallenges Conference. 2008 Oct Presented at: eChallenges Conference; 22-24.10.2008; Stockholm, Sweden URL: http://mail.srdc.com.tr/publications/2008/9.pdf
17. Aronson SJ, Clark EH, Varugheese M, Baxter S, Babb LJ, Rehm HL. Communicating new knowledge on previously reported genetic variants. Genet Med 2012 Apr 5 [FREE Full text] [doi: 10.1038/gim.2012.19] [Medline: 22481129]
18. Gerhard GS, Carey DJ, Steele Jr GD. Electronic health records in genomic medicine. In: Ginsburg GS, Willard HF, editors. Genomic and personalized medicine, 2nd ed. US: Academic Press; 2013:287-294.
19. Institute of Medicine (IOM). In: Olson S, Beachy SH, Giammaria CF, Berger A, editors. Integrating large-scale genomic information into clinical practice: Workshop summary. USA: National Academies Press; 2012.
20. Oetting WS. Clinical genetics & human genome variation: The 2008 Human Genome Variation Society scientific meeting. Hum Mutat 2009 May;30(5):852-856. [doi: 10.1002/humu.20987] [Medline: 19260058]
21. Starren J, Williams MS, Bottinger EP. Crossing the omic chasm: A time for omic ancillary systems. JAMA 2013 Mar 27;309(12):1237-1238 [FREE Full text] [doi: 10.1001/jama.2013.1579] [Medline: 23494000]
22. Marian AJ. Medical DNA sequencing. Curr Opin Cardiol 2011 May;26(3):175-180 [FREE Full text] [doi: 10.1097/HCO.0b013e3283459857] [Medline: 21415728]

Abbreviations

BioXM: BioXM Knowledge Management Environment
CCTS: Core Component Technical Specification
CDA: Clinical Document Architecture
ClinGenKB: Clinicogenomic Knowledge Base
ClinGenWeb: Clinicogenomic Web Application
CR-SNP: clinically relevant single nucleotide polymorphism
EHR: electronic health record
EMR: electronic medical record
HCRS: Health Coding Reference Server
HL7: Health Level 7
HL7 CDA R2: Health Level 7 Clinical Document Architecture release 2
ICD: International Classification of Diseases and Health Related Problems
MHDS: Minimum Health Dataset
MoH: Republic of Turkey Ministry of Health
NHDD: National Health Data Dictionary
NHIS-T: National Health Information System of Turkey
OR: odds ratio
PHR: personal health record
rsID: reference single nucleotide polymorphism identifier
SNP: single nucleotide polymorphism
TDS: Transmission Dataset
UN/CEFACT: United Nations Centre for Trade Facilitation and Electronic Business
V3: version 3
WES: whole exome sequencing
WGS: whole genome sequencing
XML: extensible markup language

Edited by G Eysenbach; submitted 25.05.14; peer-reviewed by W Hammond, A James; comments to author 08.07.14; revised version received 21.07.14; accepted 21.07.14; published 11.08.14.

Please cite as:
Beyan T, Aydın Son Y
Incorporation of Personal Single Nucleotide Polymorphism (SNP) Data into a National Level Electronic Health Record for Disease Risk Assessment, Part 2: The Incorporation of SNP into the National Health Information System of Turkey
JMIR Med Inform 2014;2(2):e17
URL: http://medinform.jmir.org/2014/2/e17/
doi:10.2196/medinform.3555
PMID:25599817

Original Paper

Incorporation of Personal Single Nucleotide Polymorphism (SNP) Data into a National Level Electronic Health Record for Disease Risk Assessment, Part 3: An Evaluation of SNP Incorporated National Health Information System of Turkey for Prostate Cancer

Timur Beyan[1], MD, PhD; Yeşim Aydın Son[1], MD, PhD

Informatics Institute, Department of Health Informatics, Middle East Technical University, Ankara, Turkey

Corresponding Author:
Yeşim Aydın Son, MD, PhD
Informatics Institute
Department of Health Informatics
Middle East Technical University
Üniversiteler Mahallesi Dumlupınar Bulvarı No:1
ODTÜ Enformatik Enstitüsü B-207, Çankaya
Ankara, 06800
Turkey
Phone: 90 312 210 7708
Fax: 90 312 210 3745
Email: yesim@metu.edu.tr

Abstract

Background: A personalized medicine approach provides opportunities for predictive and preventive medicine. Using genomic, clinical, environmental, and behavioral data, the tracking and management of individual wellness is possible. A prolific way to carry this personalized approach into routine practices can be accomplished by integrating clinical interpretations of genomic variations into electronic medical records (EMRs)/electronic health records (EHRs). Today, various central EHR infrastructures have been constituted in many countries of the world, including Turkey.

Objective: As an initial attempt to develop a sophisticated infrastructure, we have concentrated on incorporating the personal single nucleotide polymorphism (SNP) data into the National Health Information System of Turkey (NHIS-T) for disease risk assessment, and evaluated the performance of various predictive models for prostate cancer cases. We present our work as a three part miniseries: (1) an overview of requirements, (2) the incorporation of SNP data into the NHIS-T, and (3) an evaluation of SNP data incorporated into the NHIS-T for prostate cancer.

Methods: In the third article of this miniseries, we have evaluated the proposed complementary capabilities (ie, knowledge base and end-user application) with real data. Before the evaluation phase, clinicogenomic associations about increased prostate cancer risk were extracted from knowledge sources, and published predictive genomic models assessing individual prostate cancer risk were collected. To evaluate complementary capabilities, we also gathered personal SNP data of four prostate cancer cases and fifteen controls. Using these data files, we compared various independent and model-based, prostate cancer risk assessment approaches.

Results: Through the extraction and selection processes of SNP-prostate cancer risk associations, we collected 209 independent associations for increased risk of prostate cancer from the studied knowledge sources. Also, we gathered six cumulative models and two probabilistic models. Cumulative models and assessment of independent associations did not have impressive results. There was one of the probabilistic, model-based interpretation that was successful compared to the others. In envirobehavioral and clinical evaluations, we found that some of the comorbidities, especially, would be useful to evaluate disease risk. Even though we had a very limited dataset, a comparison of performances of different disease models and their implementation with real data as use case scenarios helped us to gain deeper insight into the proposed architecture.

Conclusions: In order to benefit from genomic variation data, existing EHR/EMR systems must be constructed with the capability of tracking and monitoring all aspects of personal health status (genomic, clinical, environmental, etc) in 24/7 situations, and also with the capability of suggesting evidence-based recommendations. A national-level, accredited knowledge base is a top requirement for improved end-user systems interpreting these parameters. Finally, categorization using similar, individual characteristics (SNP

patterns, exposure history, etc) may be an effective way to predict disease risks, but this approach needs to be concretized and supported with new studies.

(JMIR Med Inform 2014;2(2):e21) doi:10.2196/medinform.3560

KEYWORDS

health information systems; clinical decision support systems; disease risk model; electronic health record; epigenetics; personalized medicine; single nucleotide polymorphism

Introduction

In this miniseries, we share our work that aims to incorporate the personal single nucleotide polymorphism (SNP) data into a national level electronic health record, for example, the National Health Information System of Turkey (NHIS-T) for disease risk assessment based on genotyping information of patients.

First the literature review for SNP data incorporated electronic medical record (EMR)/electronic health record (EHR)s is presented. In addition, the requirements for the EMR/EHR systems in terms of the standardizations of terminologies and messaging are reviewed [1]. The need for a structured knowledge base, decision support approaches, systems for reporting, and risk assessment are addressed as well. Next, the NHIS-T system is overviewed, and architectural extensions to the NHIS-T for the integration of the SNP data are proposed [2]. Additionally, we have presented our design and developmental process for the complementary components of this system, for example, a knowledge base, Clinicogenomic Knowledge Base, (ClinGenKB), and end-user application, Clinicogenomic Web Application, (ClinGenWeb).

In this part, we evaluated these complementary components for prostate cancer using real, direct-to-consumer (DTC) SNP data files. We have first of all extracted and transformed clinicogenomic associations into knowledge base content, and determined assessment and reporting approaches to discern the disease risk at a personal level. Also an overall discussion of the results, limitations, and possibilities of our work covered in this miniseries is presented.

Methods

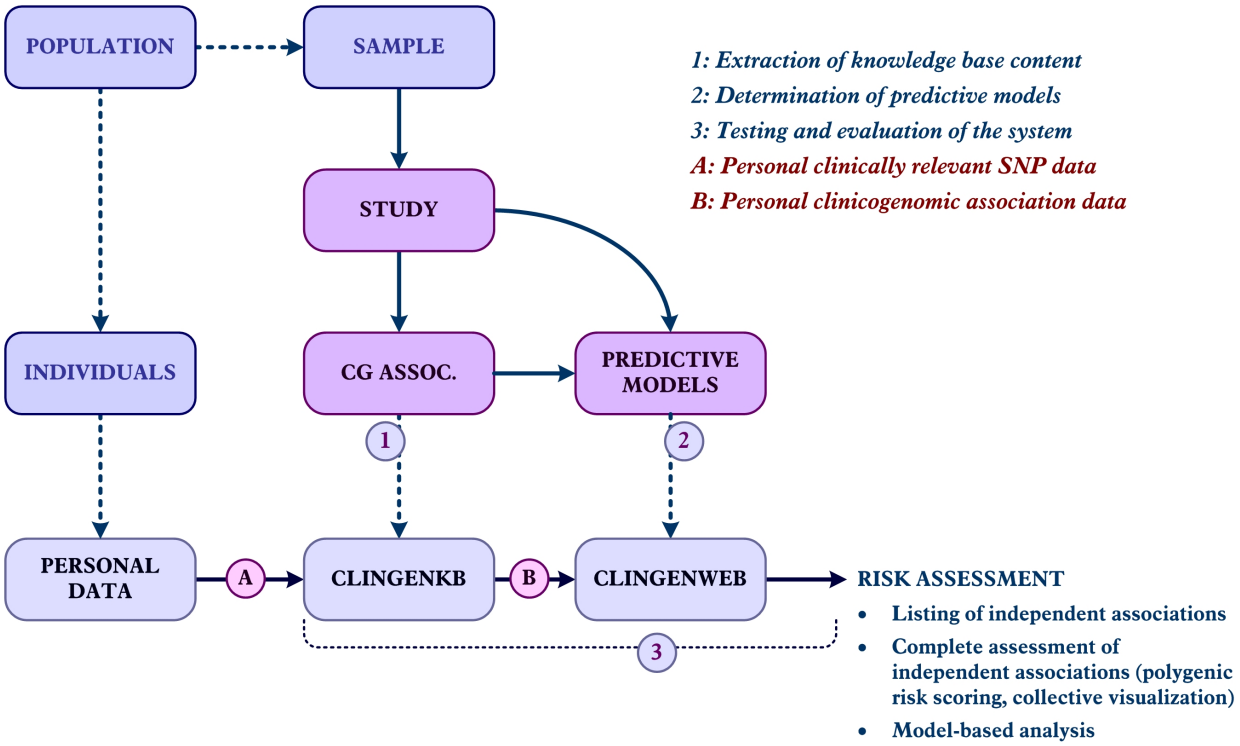
General Approach

In this article, we have focused on the evaluation of the developed ClinGenKB and ClinGenWeb for prostate cancer risk assessment.

Prostate cancer is the most common malignancy affecting men in the Western countries, it is highly heterogeneous and a multifactorial polygenic disease. The heterogeneous characteristics of prostate cancer could be partially explained by genetic factors [3]. In addition to genetic factors, age, race, family health history, endogenous hormones, diseases, environmental exposures, and various behavioral features are proposed in the literature as confounders of prostate cancer [4,5]. This complicated nature of prostate cancer, and burden on public health services, make it an ideal case to research the benefits of incorporating SNP data into an EHR for predictive, preventive, and personalized medicine approaches.

Figure 1 shows the main workflow of the process. First, the medical literature and knowledge sources to extract clinicogenomic associations between SNP alleles and increased prostate cancer risk are investigated. Additionally, the published predictive genomic models assessing individual prostate cancer risk are searched. In parallel, to evaluate our system with real data, we have gathered the personal SNP data (23andMe files) of individuals with prostate cancer and control samples. These data files are used in the evaluation phase to infer personal clinicogenomic associations based on ClinGenKB in the final stage. The independent associations and model-based prostate cancer risk assessment approaches are evaluated and compared using real personal clinicogenomic data and external data, for example, body mass index (BMI).

Figure 1. Main steps of the evaluation process. SNP=single nucletotide polymorphism; CG-ASSOC.=clinicogenomic association; CLINGENKB=clinicogenomic knowledge base; and CLINGENWEB=clinicogenomic web application.



Extraction of the Independent Clinicogenomic Associations

Collection of Associations

Since the completion of Human Genome Project, SNP-disease relationships have been extensively researched and published in the medical literature. Results of these studies are mostly collected in structured and/or narrative forms, from several clinicogenomic knowledge sources. To develop a clinicogenomic knowledge base for prostate cancer risk, we determined reliable medical sources and collected clinicogenomic associations in a standardized form.

In our study, to extract these associations, we have preferred to utilize the publicly available knowledge sources, for example, genome-wide association studies (GWAS) catalog, SNPedia, and Cancer GAMAdb. We have selected the clinicogenomic associations between SNPs and increased prostate cancer risk from these knowledge sources, excluding studies about gene-environment (eg, nutrition, drugs, chemical agents, etc) interactions. In addition, we have ignored clinicogenomic associations measuring SNP effects on the aggressivity and mortality of the prostate cancer.

As the SNP nomenclatures and notations are represented heterogeneously among different medical sources, the correct unification and standardization of identifiers had to be the initial step. We have checked all of the selected associations and matched their reference single nucleotide polymorphism identifiers (rsIDs) and alleles using Single Nucleotide Polymorphism database (dbSNP). The SNP rsIDs, which had been merged with another SNP, were updated, and allele values, which had been identified based on reverse strand, were transformed to the forward strand.

Selection of Suitable Associations

Generally, there is more than one odds ratio (OR) for every SNP-disease association in various GWAS data warehouses, depending on the diversity of studies. A selection strategy is proposed to solve these value redundancies and conflictions. For the clinicogenomic association set, we have developed a four-phased selection approach to determine a reasonable value per SNP allele.

In the first phase, because all test data was gathered from Caucasians, we have obtained the clinicogenomic association values from studies, which were performed on this race group. If there weren't any studies in the Caucasian populations, we would've preferred to use results from the mixed population as a second choice, and results from other races (Africans, Asians, etc) as the last choice. In the second phase, we have assessed the study type, for example, meta-analysis or research study, and preferred meta-analysis results. After that, if we still had more than one association value, we calculated the citation number of referenced articles. Finally, we have selected the highest OR, when needed (Table 1). With this approach, we extracted one OR value for every single SNP from the knowledge sources.

Table 1. Selection criteria for extracted associations.

Phase	Category	Order of preference
1	Race and ethnicity	Caucasians; mixed; other races (Africans, Asians, etc)
2	Study type	Meta-analysis; research study
3	Credibility of journal	Highest number of citations
4	Odds ratio	Higher number

Evidence Degree Assignment to Clinicogenomic Associations

There are still many biases and errors in the interpretation of genetic association studies. Ideally, we would prefer to evaluate association values to sort out all sources with a bias (study design, genotyping problems, publication bias, etc) of studies, but it becomes infeasible due to the time and effort needed by the professional domain experts. So, a degree of evidence quality is developed to rank all association values that are assigned.

During the extraction of clinicogenomic associations for prostate cancer, we have generated a simple approach using some indirect metrics to determine the quality of evidence degree for every association to assess the clinical utility. There are three major criteria that are used to determine the dimensions of evidence; credibility of referenced article, reliability of the study, and the scientific familiarity of SNP-disease relationships. To calculate the credibility of the referenced article, we have used the citation number of the article, the type of study, and the number of authors. Then the reliability of the study is determined based on the sample size (number of cases and controls), race, and ethnicity status are also considered. To evaluate the scientific familiarity of SNP-clinical condition relationship, we have calculated the number of the scientific articles about the SNP-prostate cancer relationship in PubMed, and the number of cumulative models, which involves the SNP allele under evaluation. These criteria are summarized in Table 2. Finally, the degree of evidence quality was calculated as the arithmetic average of all parameters for each association.

There are many SNPs reported with minor association degrees to predict prostate cancer risk. For a physician, it is impossible to interpret all disease relevant SNPs to determine the appropriate clinical action. Thus, to present an overview of the personal risk SNPs as a whole, we have categorized the magnitude of impact and the evidence degree values of associations into three classes as strong, moderate, and weak. The thresholds for the magnitude of impact (OR) were determined as strong (≥2.50), moderate (≥2.00 and <2.50), and weak (<2.00) (Table 2).

We have also extracted indirect metrics corresponding to Venice criteria to assign an evidence degree using PubMed publications and our knowledge sources (Figure 2 shows this image). This method has a potential for an automated evidence value assignment, but needs to be validated in a separate study.

Table 2. Evidence degree assignment criteria for clinicogenomic associations.

Order of preference		Value
Credibility of referenced article		
	Citation number of article	(1-15)=1, (16-50)=2, and (>50)=3
	Type of study and number of authors	(Research article and author number <10) =1; (research article and ≤10 author number <35)=2; (research article and ≤35 author number)=3; (meta-analysis and author number<7) =2; and (meta-analysis and ≥7 author number)=3
Reliability of study		
	Race and ethnicity of studies population	Other races (Africans, Asians, etc)=1; mixed=2; and Caucasians=3
	Sample size (each of case and controls)	(<100)=1; (≥100 and <1000)=2; and (>1000)=3
Scientific familiarity of SNP-disease relationship		
	Number of article for SNP-prostate cancer relationship in PubMed	(<7)=1; (≥7 and <19)=2; and (≥20)=3
	Number of cumulative models which involve SNP allele	None=1, (<3)=2, and (≥3) =3
Degree of evidence quality		
		=Total value/6
		(<1.5)= weak; (≤1.5 and <2.3)= moderate; and (≤2.3)= strong

Figure 2. Matching of our parameters and Venice criteria. SNP=single nucleotide polymorphism.

Developed Criteria for our Study	Venice Criteria		
	Amount of Evidence (Sample Size)	Extent of Replication	Protection from Bias
Citation number of article	☑		
Type of the study (research article or meta-analysis) and number of authors		☑	☑
Race and ethnicity of studied population			☑
Sample size (each of case and controls)	☑		
Number of article for SNP-prostate cancer relationship in PubMed		☑	☑
Number of cumulative models which involve SNP allele		☑	☑

Risk Assessment and Reporting Approaches

As explained in the first part of this miniseries, there are different types of risk assessment and reporting approaches, for example, listing of clinicogenomic associations and their effects as independent associations, complete representation of these SNPs using visualization techniques, calculation of disease risk using polygenic risk scoring and model-based approaches, etc. In this study, we have focused on the different models for prostate cancer.

Because most of the genomic associations have small degrees of impact, cumulative models, which contain a few critical SNPs, have been proposed previously to predict the disease risk. We have extracted cumulative models for prostate cancer risk assessment through PubMed searches.

The rsIDs and allele values of SNPs contained in models were checked and adapted to forward genomic strands based on dbSNP entries. Models, which involve additional external parameters, such as family health history, are also collected. Finally, the reference tables for all models containing the total impact of involved parameters and corresponding risk values are generated.

Among risk assessment tools other than cumulative models, there are ongoing efforts utilizing different data mining algorithms to interpret GWAS data for building various predictive models. In order to present how these modeling approaches could be implemented in our prototype system, we also included two such examples into our study. These probabilistic models are based on the works of Yücebaş and Aydın Son to assess prostate cancer risk, and were developed through a hybrid approach combining support vector machine (SVM) and Iterative Dichotomiser 3 (ID3) decision tree (DT) based on "A Multiethnic Genome-wide Scan of Prostate Cancer" dataset from the database of Genotypes and Phenotypes (dbGaP) (study accession no., phs000306, and version 2) [6,7]. The first hybrid model (only SNP model) includes 33 SNPs and their alleles, and the accuracy, precision, and recall values of this model are 71.6%, 72.69%, and 68.96%, respectively [6,7]. The second hybrid model originally was developed by integrating genotyping and phenotyping data, and contains 28 SNPs, along with clinical features; BMI, alcohol intake, and cigarette smoking. The accuracy, precision, and recall values of this model for the integrated model are 93.81%, 96.55%, and 90.92%, respectively [6,7].

Similar to cumulative models, to prepare these hybrid models, first, we checked rsIDs and adapted allele values of contained SNPs to forward deoxyribonucleic acid strands using dbSNP. After that, we have converted the results of hybrid models as association sets. Finally, we have prepared a reference table for both of the genomic risk models containing SNP parameters.

Polygenic risk scoring is an extension of cumulative model-based approaches. Different types of polygenic prediction models were developed to combine the impact of disease associated SNP data, for example, count method, log-odds method, multiplicative model, etc. The count method is the calculation of the total count of independent genomic risk alleles. The log-odds method adds together the natural logarithm of the allelic OR for each risk allele [8]. DTC testing companies typically employ a multiplicative model to calculate life time risk in the absence of an established method for combining SNP risk estimates, that is multiplication of OR of each genotype and average population risk [9].

Preparation of Test Data

To evaluate the ClinGenKB and the ClinGenWeb platforms as a part of our use case scenario, we have gathered real data (23andMe files) from the Personal Genome Project [10]. In this publicly available resource, genomic, environmental, and human trait data are integrated together. There were four 23andMe files that belonged to men who have been diagnosed with prostate cancer. All of these patients were Caucasian men, over 60 years of age. To build a demographically matched control set, we have selected all the Caucasian men older than 60 years of age as control samples. Through the Personal Genomes Project's website, we have acquired 23andMe files of 15 individual healthy Caucasian men over the age of 60 (Table 3).

Before the evaluation of the proposed workflow and the framework, first, a personal clinically relevant SNP (CR-SNP) data file for prostate cancer patients from their original 23andMe files are generated based on the clinicogenomic associations. Then the clinicogenomic associations and these test data are transferred into the ClinGenKB. Personal clinicogenomic associations were acquired by processing the personal CR-SNP data with a smart query based on the ClinGenKB. After that, acquired clinicogenomic associations were transferred into the ClinGenWeb. Also, some relevant personal health data were transferred from the Personal Genome Project website to the ClinGenWeb to be used in the interpretation of disease risks based on the models. Finally, the validity of implemented models and approaches are compared and discussed.

Table 3. Characteristics of genomic data owners.

Participant	Prostate cancer	Ancestral origin	Birth year
01-hu1213DA	Yes	Germany-Norway	1937
03-huD889CC	Yes	Ireland	1938
07-hu28F39C	Yes	United States	1943
13-hu6ED94A	Yes	United States-Austria	1950
02-hu59141C	No	United States-Canada	1937
04-huF7E042	No	United States-United Kingdom	1939
05-hu75BE2C	No	United States	1939
06-hu56B3B6	No	United States	1941
08-huB59C05	No	United States-Ireland	1943
10-hu7A2F1D	No	United States-Germany	1947
12-huD57BBF	No	United States	1949
14-huD7960A	No	Hungary-Ukraine-Russia	1951
15-hu2E413D	No	United States	1952
16-hu76CAA5	No	United States	1952
17-huA720D3	No	United States-United Kingdom	1953
18-hu63DA55	No	United States	1953
19-hu43860C	No	United Kingdom-Hungary	1954
20-huD00199	No	Germany-Poland	1954
21-huAC827A	No	United States-Sweden	1954

Evaluation of Test Data

In prostate cancer, known relevant SNPs mostly have a modest OR. Therefore, in the evaluation phase, we have assessed the total impact of the independent relevant associations based on four approaches, that is the number of SNPs based on the dominant model, the number of SNPs based on the additive model, the evidence-impact-SNP degree based on the dominant model, and the evidence-impact-SNP degree based on the additive model. The "number of SNPs", are calculated as the total count of existing relevant SNPs. In the dominant type of this model, only the count of relevant SNPs is considered, but in the additive type, the impact of homozygote SNPs is weighted twice as much compared to heterozygote SNPs. In the evidence-impact-SNP approach, for every existing SNP, we have calculated an impact degree using the evidence degrees (1, 2, and 3) and the impact degrees (1, 2, and 3). Also, similar to the number of SNPs calculated, for the additive type we have assigned 1 and 2 to heterozygote and homozygote SNPs as weighting coefficients, respectively.

After that, all the cases and controls are evaluated based on the predictive cumulative and probabilistic models. Then, results for all of the cases and controls were interpreted and compared. In the second hybrid model, where associations are based on both genotyping and clinical data, SNPs and BMI, smoking and alcohol consumption data are used. Here, due to a lack of the clinical data, the risk for some individuals could not be assessed.

In addition to the genetic factors, there are various comorbidities, sociodemographic characteristics, and environmental and

behavioral exposures that are proposed as confounders of the prostate cancer (Table 4). Therefore, we have analyzed the personal, clinical, and the environmental characteristics, which are meaningful for the prostate cancer pathogenesis, as the last step of our evaluation.

Table 4. Example list of several risk and protective factors for the prostate cancer [2,3].

Risk category	Parameters
Sociodemographic data	Age, family health history, ethnicity, and race.
Environmental sources	Nutrition and diet (animal fat, fruits, legumes, yellow-orange and cruciferous vegetables, soy foods, dairy products, fatty fish, alcohol, coffee, green tea, modified citrus pectin, and pomegranate).
	Supplements (multivitamins, vitamin E -with or without selenium, folic acid, zinc, calcium, vitamin D, retinoid, and zyflamend).
	Drugs (5 alpha-reductase inhibitors, nonsteroidal antiinflammatory drugs, statins, and toremifene).
	Medical procedures (vasectomy, barium enema, hip or pelvis x-rays, and external beam radiation therapy for rectal cancer).
	Tobacco use (tobacco products, smoking).
Personal health status (internal environment)	Medical conditions (prostatitis, prostatic intraepithelial neoplasia, syphilis, skin basal cell carcinoma, and benign prostate hyperplasia).
	Anatomic measurements (high body mass index).

Results

Independent Associations for Prostate Cancer

Initially, we have determined 87 SNP alleles from the GWAS catalog, 32 SNP alleles from the SNPedia, and 236 SNP alleles from the Cancer GAMAdb, which are all associated with increased prostate cancer risk. Through the extraction and selection processes of SNP-prostate cancer risk associations, we have excluded redundant, conflicting, and incomplete associations. Finally, a total of 209 independent associations for increased risk of prostate cancer from the studied knowledge sources were acquired. Next, the evidence and the impact categories are assigned to these associations (see Multimedia Appendix 1). The overall assessment of all these different types of clinicogenomic associations is summarized in Table 5.

Table 5. Distribution of clinicogenomic associations.

	Evidence degree			
Impact degree	Strong	Moderate	Weak	Total
Strong	0	5	2	7
Moderate	0	3	1	4
Weak	42	123	33	198
Total	42	131	36	209

Cumulative Models for Prostate Cancer

Cumulative models are the combination of the impact of several clinicogenomic associations using arithmetic operators. For some SNPs, only homozygote alleles are involved in the models (recessive model), but mostly heterozygote SNPs (dominant model) are part of the cumulative models. Both in dominant and recessive models, the values of risk SNPs are accepted as one unit of impact. Alterations of SNPs' impact values regarding homozygote and heterozygote alleles are defined as an additive model. The dominant and recessive models as examples of the cumulative predictive models retrieved from the scientific literature, and the SNP alleles included in each of the cumulative models are listed in Table 6.

In addition to Table 6, three of these cumulative models (17-SNP_Helfand, 5-SNP_Zheng and 5-SNP_Salinas) were enhanced using family health history as an additional parameter and combined SNP-family health history models were produced [11-13].

In the cumulative models, the existence of each association contributes to the total score. For example, in the 5-SNP_Zheng model, there are five different SNPs. The genetic model is dominant for three SNPs (rs1447295-A, rs16901979-A, and rs6983267-G) and recessive for the others (rs1859962-G, rs4430796-A). For dominant models, homozygote and heterozygote combinations of alleles are identified as a risk factor in the same degree. For recessive models, only homozygote combinations are considered as risk factors, whereas heterozygote combinations are accepted as harmless. Through analysis of a patient's genotype, the total impact values of clinicogenomic associations are determined and calculated additively. Besides the SNP associations, the existence of prostate cancer in family health history can be included as an additional impact factor.

The reference table for 5-SNP_Zheng model is presented as an example in Table 7. If patients without family health history have only one impact factor, the risk of having prostate cancer increases by 1.5, compared to those who have none of the impact factors. If a patient has all five risk SNPs with specified alleles, and a positive family health history for prostate cancer, the total impact is calculated to be 6. According to Table 7, this would correspond to an increased risk of 9.46 for having prostate cancer when compared to the general population. Full reference tables for all cumulative models are provided in the Multimedia Appendix 2 [11-16].

Table 6. Examples of cumulative risk prediction models for prostate cancer.

	17-SNP_Helfand [11]	9-SNP_Helfand [14]	5-SNP_Zheng [12]	5-SNP_Salinas [13]	4-SNP_Nam [15]	3-SNP_Beuten [16]
rsIDs and risk allele						
rs1819698-T						Dominant
rs2710646-A		Recessive				
rs721048-A	Recessive					
rs10934853-A	Dominant					
rs2736098-A	Recessive					
rs401681-C	Dominant					
rs1800629-A					Dominant	
rs2348763-A					Recessive	
rs1447295-A	Dominant	Dominant	Dominant	Dominant	Dominant	
rs16901979-A	Dominant	Dominant	Dominant			
rs16902094-G	Dominant					
rs445114-T	Dominant					
rs6983267-G	Dominant	Dominant	Dominant	Dominant		
rs6983561-C				Dominant		
rs10993994-T	Recessive	Recessive				
rs10896450-G	Dominant	Dominant				
rs11228565-A	Dominant					
rs12439137-G						Dominant
rs2470152-T						Dominant
rs11649743-G	Recessive					
rs1859962-G	Recessive	Recessive	Recessive	Recessive	Recessive	
rs4430796-A	Dominant	Dominant	Recessive	Recessive		
rs8102476-C	Dominant					
rs5945572-A	Dominant	Dominant				

Table 7. Reference table for 5-SNP_Zheng model.

Total impact	Odds ratio (95% CI), without FHH[a]	Odds ratio (95% CI), with FHH[a]
0	1.00 (by definition)	1.00 (by definition)
1	1.50 (1.18-1.92)	1.62 (1.27-2.08)
2	1.96 (1.54-2.49)	2.07 (1.62-2.64)
3	2.21 (1.70-2.89)	2.71 (2.08-3.53)
4	4.47 (2.93-6.80)	4.76 (3.31-6.84)
5	4.47 (2.93-6.80)	9.46 (3.62-24.72)
6	-	9.46 (3.62-24.72)

[a] FHH = family health history

Probabilistic Models for Prostate Cancer

In this study, we used two types of probabilistic models from Yücebaş and Aydın Son based on a hybrid (SVM+ID3 DT) approach; namely, first (only SNP) and second (SNP-Environmental Combined) [6,7]. When the first hybrid model (only SNP model) from Yücebaş and Aydın Son is interpreted, we have captured 154 different association sets containing the combination of several different SNPs and alleles [6]. In the second genotype-phenotype integrated model, we acquired 23 association sets containing 28 SNPs and their alleles along with BMI, smoking, and alcohol usage [7]. The complete associations of the hybrid models are listed in Multimedia Appendices 3 and 4.

In these probabilistic models, if an individual accounted for all parameters on one branch (ie, an association set), this individual has a prostate cancer risk with the accuracy, precision, and recall values of total model as presented in references [6,7]. Table 8 presents an example of the reference table for association sets of the genotype-only hybrid model.

Table 8. Reference table for the probabilistic only SNP model.

Branch_id	Total count of SNPs
Branch_ 1	4
Branch_ 2	4
Branch_ 3	7
….	….
Branch_ 154	2

Evaluation Results for Test Data

Overview

In the evaluation phase, we have studied four cases and 15 controls, which consisted of Caucasian men, age 60 years or older, and regarding independent clinicogenomic associations and risk prediction models.

Complete results of test and evaluation processes (independent association assessment, model-based evaluation, and clinical and environmental evaluation) are provided in Multimedia Appendix 5.

Results for Independent Associations

In prostate cancer, known relevant SNPs mostly have a modest OR. Therefore, in the evaluation phase, we have assessed the total impact of the independent relevant associations based on four approaches, that is the number of SNPs based on the dominant model, the number of SNPs based on the additive model, the evidence-impact-SNP degree based on the dominant model, and the evidence-impact-SNP degree based on the additive model.

The comparative evaluation results of individual clinically relevant SNPs of case and control groups regarding categorical distribution of evidence quality and impact degrees are in Multimedia Appendix 6. In these approaches, case groups were divided into two or three different subsets (three patients with high values and one patient with a low value for the dominant models, and two patients with high, one patient with moderate, and one patient with low values in terms of additive models). In control groups, there were, in particular, five people (21-huAC827A, 15-hu2E413D, 08-huB59C05, 17-huA720D3, and 06-hu56B3B6) with values higher than all cases observed. However, it must be remembered that, in the complete assessment of all SNPs, due to the remarkable number of relevant SNPs that were not analyzed, the results might be distorted.

Results for Cumulative Models

Due to a lack of family health history data of individuals, we couldn't use this data to calculate cumulative risks. In our limited number of cases, cumulative models did not have meaningful results. But, similar to the complete evaluation of independent associations, it must be considered that, nonanalyzed SNPs could be distorting the results. Results of these cumulative models are summarized in Table 9.

Table 9. Summarized results for cumulative models.

	Case			Control		
	Odds ratio≥2.5	Odds ratio<2.5	Unknown	Odds ratio≥2.5	Odds ratio<2.5	Unknown
17-SNP_Helfand	1	-	3	2[a]	10	3
9-SNP_Helfand	1	3[b]	-	1[c]	12	2
5-SNP_Zheng	-	4	-	-	15	-
5-SNP_Salinas	-	4	-	-	15	-
4-SNP_Nam	-	4	-	-	15	-
3-SNP_Beuten	-	2	2	-	13	2

[a] 02-hu59141C, 12-huD57BBF

[b] 01-hu1213DA, 03-huD889CC, and 07-hu28F39C

[c] 17-huA720D39

Results for Probabilistic Models

Regarding the probabilistic model-based interpretations; the only SNP model from Yücebaş and Aydın Son [6] wasn't successful in terms of predicting the cases. In the second model [7], where genotype and phenotype data were integrated, one patient was determined as being under risk, two patients couldn't be evaluated because of data incompleteness (smoking and alcohol consumption data), and one patient (03-huD889CC) was determined as being risk free. In control samples, only one individual (04-huF7E042) was determined as being in a risk group, but six individuals were determined as being risk free. There were eight individuals of this group that couldn't be evaluated due to data incompleteness. Although this model was produced for those of African American descent, and even though we had a limited number of cases and controls for the evaluation process, it was still the most successful approach when compared to the others. Interestingly, a patient (03-huD889CC) was determined as the risk free, and this patient was also determined as being in a low risk group according to complete assessment approaches.

Clinical and Envirobehavioral Evaluation

Prostate cancer is a polygenic multifactorial disease, and both environmental and genetic factors take important roles in its pathogenic mechanism. Therefore, if we analyze the genomic risks with clinical and environmental characteristics, we can infer more accurate results. Characteristics of cases and controls regarding clinical and environmental risk factors for prostate cancer are summarized in Table 10.

In envirobehavioral and clinical evaluation, it was found that patient "03-huD889CC" had previously been diagnosed with syphilis. In prior publications, syphilis has been reported as a risk factor for prostate cancer [2]. The healthy individuals, who had a higher risk than controls, namely "06-hu56B3B6", had basal cell carcinoma, and "21-huAC827A" had hypogonadism, that is a low level of testosterone. And both of these clinical conditions are known to decrease the prostate cancer risk [1,2]. Also, "06-hu56B3B6" and "17-huA720D3" used several risky, protective drugs and supplements regarding prostate cancer risk. In patients "08-huB59C05" and "15-hu2E413D", we did not have enough data to evaluate the risk and protective factors. In the health records of some cases and controls, there was some data about nutritional status, physical activity, and usages of supplements data, etc. But, all this data wasn't useful during the evaluation due to a lack of precise measurement information (eg, amount, period, duration, etc).

Table 10. Clinical and environmental risk factors of cases and control.

Group	Individuals	Risk factors	Protective factors
Case	01-hu1213DA	Hypercholesterolemia, BPH[a]	
Case	03-huD889CC	Syphilis	
Case	07-hu28F39C	Hypercholesterolemia, BPH[a], and lipitor	
Case	13-hu6ED94A	Obesity, hypercholesterolemia, and simvastatin	
Control	02-hu59141C	Obesity, multivitamins	T2DM[b], vegetable servings, and regular physical activity
Control	04-huF7E042	BPH[a]	TURP[c]
Control	05-hu75BE2C		Regular physical activity
Control	06-hu56B3B6	Obesity, hypercholesterolemia, chlamydia infection, alcoholism, ibuprofen, multivitamin, folic acid, vitamin E, and selenium	Basal cell skin cancer, lycopene, and pomegranate
Control	08-huB59C05	Obesity	
Control	10-hu7A2F1D	Hypercholesterolemia, atorvastatin	Nonmelanoma skin cancer, regular physical activity
Control	12-huD57BBF	Hypercholesterolemia, BPH[a] Simvastatin, aspirin, and vasectomy	Regular physical activity
Control	14-huD7960A	Overweight, hypercholesterolemia, and BPH[a]	T2DM[b]
Control	15-hu2E413D	Overweight	
Control	16-hu76CAA5	Overweight, aspirin	Omega-3 fish oil
Control	17-huA720D3	Hypercholesterolemia, aspirin, and multivitamin	Phytosterols, omega-3 fish oil, and melatonin
Control	18-hu63DA55		Omega-3 fish oil
Control	19-hu43860C	Overweight, hypercholesterolemia, and lovastatin	Nonmelanoma skin cancer
Control	20-huD00199	Overweight, hypercholesterolemia, and atorvastatin	
Control	21-huAC827A	Overweight, hypercholesterolemia, and simvastatin	Hypogonadism

[a] BPH = benign prostate hyperplasia

[b] T2DM = type II diabetes mellitus

[c] TURP = transurethral resection of the prostate

Discussion

Principal Results

In this study, we have extended the current architecture of a centralized national EHR, NHIS-T, and developed two complementary capabilities, a knowledge base (ClinGenKB) and a reporting application (ClinGenWeb), to predict the risk of diseases using SNP data.

With respect to interoperability, Health Level 7 Clinicogenomic Work Group (HL7 CG-WG) develops several standards and guidelines, and tries to overcome the chasm between the genomic laboratory and the clinical practice. In comparing current and required infrastructure characteristics, and determining a few terminology standards for genome enabled messaging, we reason NHIS-T can be adapted to HL7 CG-WG.

The unique identification of SNP data is a critical issue in clinical genomics. In our system, due to simplicity and easiness, we proposed to use rsIDs and allele values for identification of SNPs. But, to avoid any inconsistencies, it is crucial to remember that, some rsIDs have been merged over time. For this reason, SNP numbers must be checked out based on the dbSNP, and transformed into current values if required. Additionally, as different genomic strand types are the preferred choice among some clinicogenomic knowledge sources and publications, the standardization of strand identification is another important point for SNP data incorporated into clinical systems.

Regarding clinical terminology, we prefer to use existing NHIS-T standards, for example, International Classification of Diseases and Related Problems, Tenth Revision (ICD-10) for disease identification. For new data types (model name, model type, etc), we produced our own specific value categories.

To store and process the huge amount of raw variant files, in our architecture, we have proposed to store the raw and/or processed genomic data in the genomic laboratory databases,

and only to share clinically relevant variant data and/or clinicogenomic association information between partners. To derive CR-SNP data from personal SNP data, we need to use a CR-SNP resource. This resource was designed as part of a national level clinicogenomic knowledge base. This knowledge base is also utilized to transform CR-SNP data to clinicogenomic associations.

As it is emphasized in the literature, one of the most critical components of the genome enabled EHRs is the development of a national level knowledge base for clinicogenomic information. This capability must be kept up to date and manually curated by domain experts. In our study, we have developed a prototype knowledge base (ClinGenKB), which includes clinicogenomic associations for prostate cancer risk prediction.

Several different approaches are proposed to define clinical impact and evidence qualities of clinicogenomic associations in various knowledge sources. But there is still a lack of structured, objective, and comprehensive methodologies for matching, selecting, and merging different studies. In our prototype, we have proposed a simple methodology, but the best methods of determining standards to calculate, limit biases, and limit faults still need to be investigated in future clinicogenomic association studies.

ClinGenWeb is a prototype for the end-user systems that provides interpretations of the clinicogenomic associations. To evaluate our system, we have used real data from the Personal Genome Project. Collected data included 23andMe data files, ages, ethnicities, ancestral origins, clinical data, and some behavioral parameters. Age and ethnicity are extensively accepted as proven risk factors for prostate cancer. All of our cases and controls were selected from Caucasian men over 60 years old. The risk for prostate cancer is 2 in 16 for men 60 through 69 years old, and 1 in 9 for men 70 years and older [17].

ClinGenWeb uses both complete and model-based interpretations for clinicogenomic associations. Independent associations may have very little importance for clinical processes alone, but in complete interpretation, we tried to interpret all relevant data as a whole. After analyzing our results, we concluded that cases and controls could be divided into two or three different risk groups as a result of genetic heterogeneity. With the commissioning of whole genome sequencing/whole exome sequencing (WGS/WES) in clinical practice, similarity measurements of clinically relevant SNP patterns may be a new way of producing predictive models in genomic medicine, but this approach needs to be supported with more phenotypic data, and needs to be tested in larger study samples.

There are several cumulative models proposed to predict prostate cancer, but we couldn't acquire meaningful results with these models in our subjects. Another original approach was to use the probabilistic (SVM+ID3 DT) model-based associations. However, the only SNP model of this approach was not successful, but the second model, which integrates genotype and clinical data, was partly consistent. Unfortunately, the number of available holistic envirogenomic models that could be implemented here is limited. The probabilistic model utilized was produced for men of African American, Latin, and Japanese descent, and we have used the submodel template generated for African American individuals, as their genetic background is expected to carry a higher number of common SNPs with the Caucasian population than Latin or Japanese populations.

Another critical point is that clinical, environmental, and behavioral data can be used to explain pathogenic and clinical heterogeneity, and to clarify the complexity of results. With the support of clinical and behavioral data, we could interpret some contradictory results. Because, most of the environmental and behavioral data wasn't stored in EMR/EHRs in a structured manner, we generated the functionality to add these types of data at the end-user level.

Due both to the bipartite structure of our interpretations (ie, conversion of CR-SNP into clinicogenomic associations and final clinical interpretations of associations), and the fact that the final interpretation was accomplished at the end-user side, we combined both clinicogenomic associations and external parameters (such as BMI), which have been recorded or tracked by end-users to support the decision making.

Limitations

Complete implementation of SNP data incorporated NHIS-T in real systems was not possible due to the regulative and the technical issues at this stage. So, we restricted our focus to develop complementary capabilities as prototypes for NHIS-T, namely, the ClinGenKB and the ClinGenWeb, which specifically targeted prostate cancer risk prediction.

GWAS research is based on the "common disease, common variant" hypothesis. However, some authors proposed that common variants can explain only a modest part of complex diseases and so the "common disease, rare variant(s)" hypothesis was recently put forward [18]. Clinicogenomic associations used to build the knowledge base in this study are based on recent developments in the GWAS research and literature. In our study, we have only used SNP data, but recent studies show that different variants (Copy Number Variations, etc) are also responsible for clinical conditions.

Also in the ClinGenKB, our critical focus was to generate a structured clinicogenomic representation for only risk prediction for prostate cancer. But, in the literature, there are several kinds of information related to different stages of clinical decision processes, for example, prognosis, pharmacogenomic, etc. In the real world project, this prototype has to be enhanced with additional types of associations and diseases.

We obtained case and control data from the Personal Genome Project to evaluate our system, but the number of cases and controls were so limited. To determine the value of this system in clinical settings, more comprehensive data on genomic, environmental, family health, and clinical conditions are needed. Unfortunately, none of the cases and controls had family health history data, and we couldn't involve this critical parameter in our evaluation processes. Existing clinical data about subjects didn't reflect the clinical and pathological heterogeneity of the prostate cancer. In particular, we did not have precise measurement information (amount, period, duration, etc) about behavioral characteristics of subjects (diet, physical activity,

supplements, etc), and we couldn't interpret the possible effects of these parameters on prostate cancer risk.

Another limitation was in aligning the terminologies of the clinical and bioinformatical domains in a consistent way. ICD classification is accepted as a standard for disease classification in many countries including Turkey. But ICD-10 is not useful to manage all levels of clinical, pathologic, and genetic heterogeneities. It is expected that it will be managed in the next version, ICD-11 that will be released in 2015 and the new release can be integrated with other medical terminologies such as Systematized Nomenclature of Medicine Clinical Term (SNOMED-CT) [19]. Nevertheless, as proposed earlier, it is an unavoidable requirement to develop a new taxonomy of diseases, which will be based on information commons and a knowledge network, combining molecular data, social data, environmental data, clinical data, and health outcomes [20].

In the current study, due to the ethnic characteristics of our subjects, we have primarily preferred the studies performed with Caucasians to collect the clinicogenomic associations from the literature. But, the terms of ethnicity and race are sociocultural constructs affected by both biological and environmental factors. For this reason, for a real world NHIS-T system, genotyping data from the Turkish population is needed to build the working knowledge base.

Also, predictive models that will be used in clinical settings need to be validated. Especially, we need approaches to assess the complete analysis of clinically relevant SNPs. With the commissioning of WGS/WES in the clinical practice, similarity measurements of clinically relevant SNP patterns may be a new way to produce predictive models in genomic medicine, but this approach needs to be enhanced with further phenotypic data, and to be tested in large study samples.

On the other hand, the number of available holistic envirogenomic models is limited. As most of the complex diseases are progressing as an interaction of genomic and environmental factors, more envirogenomic data also need to be developed to build predictive disease models.

Comparison With Prior Work

GeneInsight Suite is an impressive application environment to evaluate and share sequencing based test results. GeneInsight Clinic can be integrated with EMRs or can be used as a standalone system. It manages knowledge, and facilitates reporting. GeneInsight Network (VariantWire) provides the mechanism to connect laboratories and providers. Interpretations of sequencing based tests are shared with corresponding caregiver organizations using this system. GeneInsight Suite allows clinicians to receive updates when new information on previously unknown variants is certified for clinical use.

There are critical differences between the proposed system and GeneInsight. First, our system is designed as part of a central national level EHR. In the United States, the architecture of EHR systems is more federated. Both systems include a knowledge base and applications for the end-users.

In GeneInsight, the interpretation and reinterpretation of critical variants are reported for clinical use. These interpretations do not involve external data, which is not included in the EMR. But, in the proposed system, the clinical interpretation of SNP data is divided into two sequential processes, that is the conversion of CR-SNP into clinicogenomic associations, and the clinical interpretation of them. Final interpretation is completed at the end-user application, and so it is possible to use additional data for the risk prediction (environmental, behavioral, etc). These processes are finalized based on predictive models and automated analysis techniques.

Conclusions

Today, the health care systems are continuously evolving and transforming under the influence of developments in technology and globalization. A revolutionary paradigm shift is changing the focus of medicine from the traditional provider-centric approach to patient-centric personalized medicine. This paradigm shift, dramatically transforms clinical processes, medical education, and research in theory and practice. The commissioning of new health services based on emerging technologies (mobile health systems, pervasive applications, environmental sensors, body area sensor networks, etc) also dramatically supports these emerging trends.

But in the light of the literature on personalized medicine, we can argue that the area of biomedical informatics has not begun to show its major effect on health care systems, and the major shifting in health care practices is expected soon via genomic technologies. When we look at the big picture, we can see the emergence of evidence-based managed health care systems with knowledge discovery capabilities driven by big data and knowledge infrastructures for sustainable, fair, and effective care services.

In this respect, we consider that the next generation of health information systems will be constructed based on tracking and monitoring all aspects of individual health status through 24/7, and implementing evidence-based recommendations to empower individuals. Today, most of the personal, behavioral, and environmental data is not a subject of EMR/EHR, or even PHR contents. Characteristics of most environmental and behavioral data require frequent measurements and (nearly) continuous tracking. And, possibly if we extend PHR content (with genomic data) toward involving environmental and behavioral factors, we can add value to disease risk assessment and prediction.

As we emphasized before, a national level manually curated and accredited knowledge base is the most important component of evidence-based decision making. Based on this knowledge base, collected risk data will gain a predictive meaning, and any new discovery in clinical sciences will be reflected for individuals by the reinterpretation of collected data. At this point, we need additional and improved analytic tools based on genomic and environmental parameters. We aim to develop a knowledge repository integrating some knowledge bases with semantic technologies, and adding some automatic evaluation techniques to make it easier to extract and manually curate existing references for the domain experts.

Regarding the challenges facing health care systems, along with the effective provision of public health services and associated financial burdens, most of the important diseases are of a

complex nature. In the pathogenesis of complex diseases, the interaction of genetic and environmental factors has critical importance, and ethnicity, race, and geographic factors may play distinctive roles. Hence, it is necessary to have the appropriate clinicogenomic information about the target population. Clinical data, environmental factors, and family health history are critical components, and there is a need to study the relationships between these parameters and genomic factors. Eventually, it will be required both to conduct envirogenetic studies in order to acquire original data for population, and to enhance the NHIS-T data model for collecting these types of data.

The omics area is not only represented by genomic data, and in the near future different types of omics data will be available for the routine clinical practices, for example, transcriptomics, proteomics, metabolomics, and epigenomics. Also, systems medicine offers possibilities that will increase the effectiveness of risk prediction strategies.

In addition, we aim to enhance our system by integrating data warehouses for research. With this capability, integrated genomic and environmental datasets can also be used for clinical research. We will extract the meaningful relationship patterns via this system and, by using these patterns; we can calculate the risks of groups who have similar characteristics, for example, family members or communities.

The major aim of our system is to provide true and actionable information for patients and their family practitioners. Our system will process collected data and return evidence-based recommendations to the individuals to make them responsible for their preferences and consequences. The empowerment of individuals to participate in their health care decisions is an emerging trend in personalized medicine. At this point, we need more curated information sources and visual representation approaches intended for unprofessional individuals. Areas of representation and reporting of clinicogenomic results should focus on developing new approaches, techniques, and tools.

In the last 10 years, there has been a great effort to accomplish a transformation to a national health care system based on information technologies in Turkey. But yet, practical applications of personal genomics and its integration into health care services are in its infancy, and studies about personalized medicine are at the academic level.

Our architecture and prototype, which aim to incorporate personal SNP data into the NHIS-T, are also in their preliminary stage. However, we need additional vision, research, work, and tools to extend our EHR capabilities for the future genome enabled health care systems. We believe that our work will be a starting point for a predictive and preemptive personalized national health care system.

Acknowledgments

We would like to thank Biomax Informatics AG for their permit, help, and support to use BioXM Knowledge Management Environment.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Complete list of independent associations.

[XLSX File (Microsoft Excel File), 26KB - medinform_v2i2e21_app1.xlsx]

Multimedia Appendix 2

Reference tables for cumulative models.

[XLSX File (Microsoft Excel File), 10KB - medinform_v2i2e21_app2.xlsx]

Multimedia Appendix 3

List of first probabilistic model based associations (only SNP model).

[XLSX File (Microsoft Excel File), 27KB - medinform_v2i2e21_app3.xlsx]

Multimedia Appendix 4

List of second probabilistic model based associations (SNP-environmental combined).

[XLSX File (Microsoft Excel File), 11KB - medinform_v2i2e21_app4.xlsx]

Multimedia Appendix 5

Complete results of test and evaluation processes.

[XLSX File (Microsoft Excel File), 13KB - medinform_v2i2e21_app5.xlsx]

Multimedia Appendix 6

Analysis of total number and values of relevant personal SNPs.

[XLSX File (Microsoft Excel File), 11KB - medinform_v2i2e21_app6.xlsx]

References

1. Beyan T, Aydın Son Y. Incorporation of personal single nucleotide polymorphism (SNP) data into a national level electronic health record for disease risk assessment, part 1: An overview of requirements. JMIR Med Inform 2014 Jul 24;2(2):e15 [FREE Full text] [doi: 10.2196/medinform.3169]
2. Beyan T, Aydın Son Y. Incorporation of personal single nucleotide polymorphism data into a national level electronic health record for disease risk assessment, part 2: The incorporation of SNP into the NHIS-T. JMIR Med Inform 2014;2(2):e17 [FREE Full text] [doi: 10.2196/medinform.3555]
3. Boyd LK, Mao X, Lu YJ. The complexity of prostate cancer: Genomic alterations and heterogeneity. Nat Rev Urol 2012 Nov;9(11):652-664. [doi: 10.1038/nrurol.2012.185] [Medline: 23132303]
4. National Cancer Institute. Risk factors for prostate cancer development. URL: http://www.cancer.gov/cancertopics/pdq/prevention/prostate/healthprofessional/page3 [accessed 2013-11-29] [WebCite Cache ID 6LURu5Igc]
5. Sartor AO. Risk factors for prostate cancer. 2014 Jun. URL: http://www.uptodate.com/contents/risk-factors-for-prostate-cancer [accessed 2013-11-29] [WebCite Cache ID 6LUSDqPkd]
6. Yücebaş SC. METU Thesis Collection, Ankara. 2013 Sep. A hybrid feature selection model for GWAS URL: http://etd.lib.metu.edu.tr/upload/12616393/index.pdf [accessed 2013-12-07] [WebCite Cache ID 6LhGYUid7]
7. Yücebaş SC, Aydın Son Y. A prostate cancer model build by a novel SVM-ID3 hybrid feature selection method using both genotyping and phenotype data from dbGaP. PLoS One 2014;9(3):e91404 [FREE Full text] [doi: 10.1371/journal.pone.0091404] [Medline: 24651484]
8. Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet 2009 Sep 15;18(18):3525-3531 [FREE Full text] [doi: 10.1093/hmg/ddp295] [Medline: 19553258]
9. Nusbaum R, Leventhal KG, Hooker GW, Peshkin BN, Butrick M, Salehizadeh Y, et al. Translational genomic research: Protocol development and initial outcomes following SNP testing for colon cancer risk. Transl Behav Med 2013 Mar 1;3(1):17-29 [FREE Full text] [doi: 10.1007/s13142-012-0149-0] [Medline: 23565131]
10. Personal Genome Project (PGP). URL: https://my.pgp-hms.org/ [accessed 2014-07-31] [WebCite Cache ID 6RTmx3BoA]
11. Helfand BT, Kan D, Modi P, Catalona WJ. Prostate cancer risk alleles significantly improve disease detection and are associated with aggressive features in patients with a "normal" prostate specific antigen and digital rectal examination. Prostate 2011 Mar 1;71(4):394-402 [FREE Full text] [doi: 10.1002/pros.21253] [Medline: 20860009]
12. Zheng SL, Sun J, Wiklund F, Smith S, Stattin P, Li G, et al. Cumulative association of five genetic variants with prostate cancer. N Engl J Med 2008 Feb 28;358(9):910-919. [doi: 10.1056/NEJMoa075819] [Medline: 18199855]
13. Salinas CA, Koopmeiners JS, Kwon EM, FitzGerald L, Lin DW, Ostrander EA, et al. Clinical utility of five genetic variants for predicting prostate cancer risk and mortality. Prostate 2009 Mar 1;69(4):363-372 [FREE Full text] [doi: 10.1002/pros.20887] [Medline: 19058137]
14. Helfand BT, Fought AJ, Loeb S, Meeks JJ, Kan D, Catalona WJ. Genetic prostate cancer risk assessment: Common variants in 9 genomic regions are associated with cumulative risk. J Urol 2010 Aug;184(2):501-505 [FREE Full text] [doi: 10.1016/j.juro.2010.04.032] [Medline: 20620408]
15. Nam RK, Zhang WW, Trachtenberg J, Seth A, Klotz LH, Stanimirovic A, et al. Utility of incorporating genetic variants for the early detection of prostate cancer. Clin Cancer Res 2009 Mar 1;15(5):1787-1793 [FREE Full text] [doi: 10.1158/1078-0432.CCR-08-1593] [Medline: 19223501]
16. Beuten J, Gelfond JA, Franke JL, Weldon KS, Crandall AC, Johnson-Pais TL, et al. Single and multigenic analysis of the association between variants in 12 steroid hormone metabolism genes and risk of prostate cancer. Cancer Epidemiol Biomarkers Prev 2009 Jun;18(6):1869-1880 [FREE Full text] [doi: 10.1158/1055-9965.EPI-09-0076] [Medline: 19505920]
17. National Cancer Institute. Genetics of prostate cancer. URL: http://www.cancer.gov/cancertopics/pdq/genetics/prostate/healthprofessional [accessed 2013-12-03] [WebCite Cache ID 6Lb1DVtBj]
18. Lake NJ, Bozaoğlu K, Khan AW, Jowett JBM. Approaches for dissection of the genetic basis of complex disease development in humans. In: Çalışkan M, editor. Genetic diversity in microorganisms. Rijeka, Croatia: InTech; 2012:309-339.
19. Zafar A, Ezat WP S. Development of ICD 11: Changes and challenges. BMC Health Serv Res 2012;12(Suppl 1):I8. [doi: 10.1186/1472-6963-12-S1-I8]
20. Committee on a Framework for Development a New Taxonomy of Disease, National Research Council. Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease. USA: National Academies Press; 2011.

Abbreviations

BMI: body mass index
ClinGenKB: Clinicogenomic Knowledge Base
ClinGenWeb: Clinicogenomic Web Application
CR-SNP: clinically relevant single nucleotide polymorphism
dbSNP: Single Nucleotide Polymorphism database
DT: decision tree
DTC: direct-to-consumer
EHR: electronic health record
EMR: electronic medical record
GWAS: genome-wide association studies
HL7: Health Level 7
HL7 CG-WG: Health Level 7 Clinicogenomic Work Group
ICD: International Classification of Diseases and Health Related Problems
ICD-10: International Classification of Diseases and Related Problems, Tenth Revision
ID3: Iterative Dichotomiser 3
NHIS-T: National Health Information System of Turkey
OR: odds ratio
PHR: personal health record
rsIDs: reference single nucleotide polymorphism identifiers
SNP: single nucleotide polymorphism
SVM: support vector machine
WES: whole exome sequencing
WGS: whole genome sequencing

Edited by G Eysenbach; submitted 25.05.14; peer-reviewed by W Hammond; accepted 15.07.14; published 19.08.14.

Please cite as:
Beyan T, Aydın Son Y
Incorporation of Personal Single Nucleotide Polymorphism (SNP) Data into a National Level Electronic Health Record for Disease Risk Assessment, Part 3: An Evaluation of SNP Incorporated National Health Information System of Turkey for Prostate Cancer
JMIR Med Inform 2014;2(2):e21
URL: http://medinform.jmir.org/2014/2/e21/
doi:10.2196/medinform.3560
PMID:25600087

Original Paper

Barriers Over Time to Full Implementation of Health Information Exchange in the United States

Clemens Scott Kruse[1*], BS, MBA, MSIT, MHA, PhD; Verna Regier[1*], MHA; Kurt T Rheinboldt[1*], MSW, MEd

School of Health Administration, College of Allied Health Professions, Texas State University, San Marcos, TX, United States

[*]all authors contributed equally

Corresponding Author:
Clemens Scott Kruse, BS, MBA, MSIT, MHA, PhD
School of Health Administration
College of Allied Health Professions
Texas State University
HPB, 2nd Floor
601 University Dr
San Marcos, TX, 78666
United States
Phone: 1 210 355 4742
Fax: 1 512 245 8712
Email: scottkruse@txstate.edu

Abstract

Background: Although health information exchanges (HIE) have existed since their introduction by President Bush in his 2004 State of the Union Address, and despite monetary incentives earmarked in 2009 by the health information technology for economic and clinical health (HITECH) Act, adoption of HIE has been sparse in the United States. Research has been conducted to explore the concept of HIE and its benefit to patients, but viable business plans for their existence are rare, and so far, no research has been conducted on the dynamic nature of barriers over time.

Objective: The aim of this study is to map the barriers mentioned in the literature to illustrate the effect, if any, of barriers discussed with respect to the HITECH Act from 2009 to the early months of 2014.

Methods: We conducted a systematic literature review from CINAHL, PubMed, and Google Scholar. The search criteria primarily focused on studies. Each article was read by at least two of the authors, and a final set was established for evaluation (n=28).

Results: The 28 articles identified 16 barriers. Cost and efficiency/workflow were identified 15% and 13% of all instances of barriers mentioned in literature, respectively. The years 2010 and 2011 were the most plentiful years when barriers were discussed, with 75% and 69% of all barriers listed, respectively.

Conclusions: The frequency of barriers mentioned in literature demonstrates the mindfulness of users, developers, and both local and national government. The broad conclusion is that public policy masks the effects of some barriers, while revealing others. However, a deleterious effect can be inferred when the public funds are exhausted. Public policy will need to lever incentives to overcome many of the barriers such as cost and impediments to competition. Process improvement managers need to optimize the efficiency of current practices at the point of care. Developers will need to work with users to ensure tools that use HIE resources work into existing workflows.

(JMIR Med Inform 2014;2(2):e26) doi:10.2196/medinform.3625

KEYWORDS

medical informatics; electronic health record (EHR); electronic medical records (EMR); health information technology (HIT); quality improvement; national health policy; workflow; past trends

Introduction

Health Information Exchange (HIE) is not a new concept. It was prioritized in a national agenda in the United States by President Bush in 2004 [1]. Physicians understand and agree with the altruistic benefit that HIE can enable [2], but many barriers prevent its widespread adoption. Enterprise-wide savings have full implementation range of $8.1-$77.8 billion [3,4] and a pay-back period as low as 2.1 years [5], but the disjointed nature of the health system in the United States creates a disconnect between long-term savings of payers and short-term investment of providers. Many studies have examined the barriers to adoption, but no research has examined these barriers over time.

An HIE is the electronic transfer of clinical and administrative information [3], across diverse and often competing health care organizations [2], at the state or regional levels [6], delivering the right information to the right person at the right time. The use of HIE networks has the potential to reduce up to 18% of patient safety errors generally and as many as 70% of preventable adverse drug events across the care continuum [7]. The HIE concept has the potential to reduce health care costs in the United States through a reduction in unnecessary medical tests and procedures, by improving communication about patients' latest medication regimens, laboratory test results, and diagnostic procedures [7]. The HIE concept also has the potential to improve infection control practice. For example, Kho et al found that across a large metropolitan area, 286 unique patients generated 587 admissions accounting for 4335 inpatient days where the receiving hospital was not aware of the prior history of methicillin-resistant Staphylococcus aureus (MRSA) [8].

The Institute of Medicine (IOM) determined that automation of clinical data through electronic methods would result in better patient care [8]. What followed in 2004, was Executive Order 13335 which set a goal to fully adopt electronic health records (EHR) within ten years [1]. Within a short amount of time, several HIEs, under both public and private funding, appeared on the health care landscape in the United States. There was no standard for an organization that enabled the exchange, or the exchange itself. Lack of standards continues today, which enables innovation in design, but also does not help new startup initiatives start with a successful model. Studies demonstrated the advantages to the concept of health information exchange: cost, quality, safety, better patient care, fewer repeat tests, reduced readmissions, and the ability to identify "drug hoppers" [1-5]. The HIE is defined in concept, as in the previous paragraph, but not in design.

In the first few years after President Bush's Executive Order, most HIEs initiated had failed due to their own fiscal weight and the absence of a viable business plan. Barriers to adoption were listed in the literature: lack of a viable business plan to sustain the HIE and acceptance by providers and patients [6], privacy/security concerns [8,9], usability [10], lack of technical support or technology gaps [9,10], missing data [11,12], disruption of workflow [9,10,12], startup costs from public and private dollars [13,14], lack of experience in the concept of HIE, and interference with competition [14].

In 2009 the United States Congress passed the Health Information Technology for Economic and Clinical Health (HITECH) Act, as part of the American Recovery and Reinvestment Act (ARRA), which earmarked $19.2 billion as incentives for providers to adopt the EHR and to participate in HIE [1]. The National Coordinator for Health Information Technology (ONC) created the State HIE Cooperative Agreement Program which sponsored public grants specifically for the startup of HIEs and Regional Health Information Organizations (RHIOs). States were encouraged to match the federal dollars to also incentivize the HIE concept. The intent was to help new HIE initiatives overcome the initial fiscal problems until the concept of HIE was accepted and supported in the health care community. The act also enables the comingling of private and public funds because the public money was issued as a grant. Private organizations interested in the pursuit of an exchange could augment their own budget with public grant money to provide an advantageous fiscal position not previously available.

The HITECH Act promotes the electronic exchange of clinical health data across organizations with the expectation that access to comprehensive patient information will help clinical decision-making. Once again, the federal government defined what the HIE should do, but stopped short of defining how it should be done. There is general agreement that access to a patient's medical record at the point of care will help to avoid duplicative tests, increase administrative efficiency, improve disease management, and ultimately result in cost savings. Interoperable health information may also help to identify and avoid medication complications, thus increasing patient care and safety. However, the fragmented health system in the United States presents many structural, economic, and cultural challenges to achieving a robust environment of electronic data exchange [4].

In light of federal efforts to facilitate the adoption of EHRs and formation of HIEs, the ONC, under the auspices of the Department of Health and Human Services, tracks EHR adoption rates for office-based providers and hospitals on its Health IT Dashboard. In 2008, 17% of office-based providers used a basic EHR, increasing to 40% in 2012. Similarly, in 2008, 13% of hospitals implemented a basic EHR, growing to 56% in 2012. A basic EHR includes patient demographics, patient problem lists, medication histories, clinical notes, electronic orders for prescriptions, laboratory results, and imaging results [15]. In regard to the advancement of HIEs, the 2013 e-Health Initiative Survey on Health Data Exchange identified 84 data exchange initiatives out of 315 that are at advanced stages of operation and thus able to support data exchange. This represents an increase from 57 advanced initiatives in 2011. Growth trends indicate a positive relationship between EHR adoption and HIEs with exchange capacity [16].

However, it is important to note the distinction between HIE capacity to exchange data from the actual rates of data exchange by providers and health organizations. The absence of a viable business plan or standard organizational structure of the

exchange may have caused the rate of exchange to be lower than desired. A recent study identified similar growth in hospitals exchange of health information with other entities. Exchange rates to providers outside the hospital's organization were 41% in 2008 and increased to 58% in 2012. In contrast, data on HIE utilization rates among office-based providers is more limited to narrower studies that focus on specific specialties, user types, and geographic regions [17].

Although research has analyzed the EHR, HIE, and barriers to adoption of both, no study maps the barriers reported over time. This gap in the literature provides the basis for this article. The aim of this study is to examine the frequency of barriers as listed in published material from PubMed (MedLine), CINAHL, and Google Scholar. From this analysis, a data map over time is developed to better understand the dynamic nature of the results. The results of this study enable future researchers to develop empirical models and policy makers to exploit the successful levers that generate a desired result.

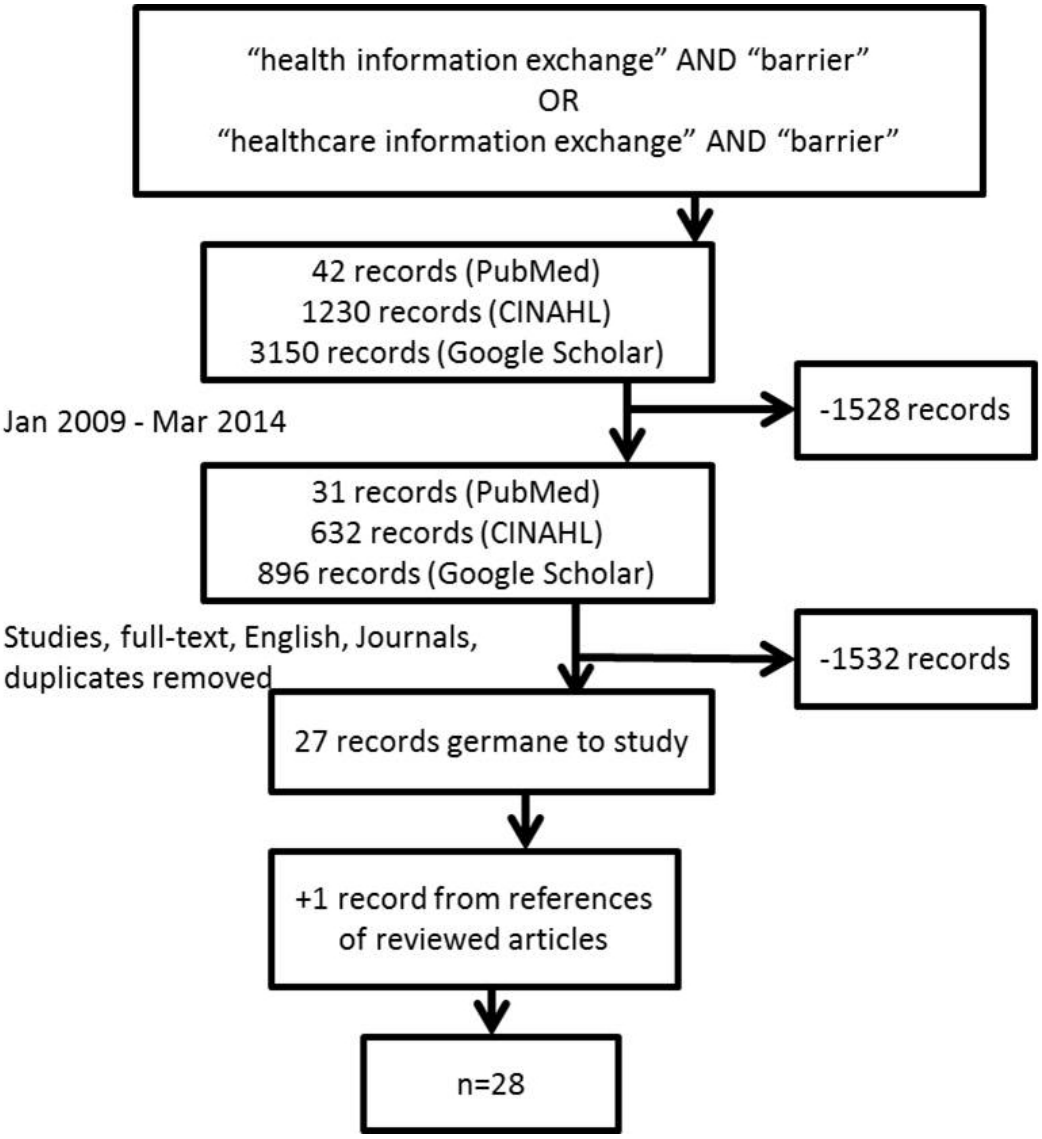
Methods

Search terms were selected based on the experience of the authors in the field of health care administration. The time-frame for the literature review of 1993-2014 was selected out of convenience. It was assumed that two decades would be sufficient to capture trends. The years under study were Jan 2009-Mar 2014. This span was chosen because of the incentives (grants) enabled by the ARRA, and also a concentrated study on these years was expected to enhance the results.

Figure 1 illustrates the literature review process that identified sources consisting of empirical studies, articles, editorials, commentaries, opinion papers, organizational theories, and text books. The window of time for this study eliminated 1528 records. Focusing on studies, full-text, English, academic journals, and eliminating duplicates resulted in the removal of an additional 1532 records. After 27 articles were identified and reviewed, one additional article was selected from the references of multiple studies. The final sample was 28.

There were no human subjects in this study; all information came from secondary data sources. The studies used in this research were sources that were publically available, and the subjects could not be identified either directly or through identifiers linked to the subject. This qualifies under "exempt" status in 45 CFR 46. Therefore, IRB review was not required, and consent from subjects was irrelevant.

Figure 1. The literature review process.



Results

Table 1 illustrates the results organized by by the year in which the literature was published. This table lists the associated number, synchronized with the references section, the journal publishing the article, the associated method, and the barriers listed. For example, Rudin R et al. identified an inequity between providers of the information and others (in disparate organizations) that benefit from the presence of the information. This was categorized into the barrier of "impedes competition" [12]. Patel V et al identified an inequity of those who pay for participation in the HIE and those who benefit, such as the patient. This was categorized as both "impedes competition" and "misaligned incentives" [13]. Numbers in the column labeled "Art#" are not in order because the article was also used in the literature review, and the authors wanted to synchronize the list of articles in this study with its references section.

The authors categorized the barriers into 16 common themes, and listed in parentheses the barrier interpreted from the studies' results. A total of 28 articles identified 16 unique barriers. Barriers are listed in the order of most identified. The numbers in each column correspond to the study itself, synchronized with the references. Three articles in Table 1 list the numbers in italics. These articles were literature reviews. The authors chose to include these studies for reasons of consistency and reliability.

The number of articles identified and reviewed from 2009-2014 were 1, 5, 7, 7, 7, and 1, respectively, while the number of barriers listed from 2009-2014 were 2, 12, 11, 8, 7, and 2, respectively. In 2009, the only barriers listed were cost and physician resistance [4]. In 2010, the 12 barriers were: cost, efficiency/workflow, lack of technical support/technology gap, impedes competition, value of HIE is difficult to measure, privacy/security, usability, heavily dependent on leadership of the organization, liability, physician resistance, decreases quality, and increases error [2,10,18-20]. In 2011, the 11 barriers listed were: cost, efficiency/workflow, lack of technical support/technology gap, impedes competition, value of HIE is difficult to measure, privacy/security, clinical data missing when needed, usability, heavily dependent on leadership of the organization, liability, misaligned incentives [12,13,21-25]. In 2012, the 8 barriers listed were: cost, efficiency/workflow, lack of technical support/technology gap, impedes competition, value

of HIE is difficult to measure, privacy/security, clinical data missing when needed, and liability [14,26-31]. In 2013, the barriers listed were cost, efficiency/workflow, impedes competition, value of HIE is difficult to measure, clinical data missing when needed, usability, heavily dependent on leadership of the organization, lack of standards, and misaligned incentives [32-38]. In 2014, the two barriers listed were efficiency/workflow and usability [39]. Table 2 organizes the barriers listed in the literature by the year in which the literature was published.

Table 1. Studies and barriers identified.

Art #	Study	Date	Barriers
4	Adler-Milstein J, Bates DW, Jha AK. Regional Health Information Organizations: progress and challenges	2009	viable business model, failure to obtain sufficient participation, cost
2	Fontaine P, Ross SE, Zink T, Schilling LM. Systematic review of health information exchange in primary care practices	2010	cost, security and privacy issues, liability, leadership, strategic planning, and competition, technical gap
7	Vest J, Gamm L. More than just a question of technology: Factors related to hospitals' adoption and implementation of health information exchange	2010	cost, competition, privacy concerns, legal liability
10	Ross SE, Schilling LM, Fernald DH, Davidson AJ, West DR. Health information exchange in small-to-medium sized family medicine practices: motivators, barriers, and potential facilitators of adoption	2010	cost, workflow, tech support, competition, non-solidarity, usability
11	Tham E, Ross SE, Mellis BK, Beaty BL, Schilling LM, Davidson AJ. Interest in health information exchange in ambulatory care: a statewide survey	2010	missing data
18	Wright A, Soran C, Jenter C. Physician attitudes toward health information exchange: results of a statewide survey	2010	privacy, difficulty to assess value of HIE
19	Dixon B, Zafar J. A framework for evaluating the costs, effort, and value of nationwide health information exchange	2010	technology gap
12	Rudin R, Volk L, Simon S, Bates D. What affects clinicians' usage of health information exchange	2011	gaps in data, workflow, usability, billing (cost), inequity between providers of information and those who benefit from the information (competition)
13	Patel V, Abramson EL, Edwards A, Malhotra S, Kaushal R. Physicians' potential use and preferences related to health information exchange	2011	costs, tech support, inequity of those who pay , and those who benefit (impedes competition and misaligned incentives), workflow, usability
20	Joshi JK. Clinical Value-Add for Health Information Exchange (HIE)	2011	quality of care, effect on patients, cost, error, organizational efficiency, acceptance by physicians and patients.
21	Korst LM, Aydin CE, Signer JM, Fink A. Hospital readiness for health information exchange: Development of metrics associated with successful collaboration for quality improvement	2011	strong leadership, tech support, value of data
22	Adler-Milstein J, Bates DW, Jha AK. A survey of health information exchange organizations in the United States: implications for meaningful use	2011	cost, leadership, lack of value
23	Lluch M. Health care professionals' organisational barriers to health information technologies-A literature review	2011	structure of health care organizations (ownership), tasks (workflow), people policies (liability), incentives (cost), information and decision processes (tech support)
24	Gadd CS, Ho YX, Cala CM, Blakemore D, Chen Q, Frisse M, Johnson K. User perspectives on the usability of a regional health information exchange	2011	not user-friendly (efficiency), need additional tech support, data incomplete (data missing when needed)
25	Hincapie AL, Warholak TL, Murcko AC, Slack M, Malone DC. Physicians' opinions of a health information exchange	2011	lack of value, technology gaps, gaps in data
14	Pevnick J, Claver M, Dobalian A, Asch S, Stutman H, Tomines A, Fu P. Provider stakeholders' perceived benefit from a nascent health information exchange: A qualitative analysis	2012	legal concerns (liability), data security, costs, competition, bureaucracy (efficiency)
26	Williams C, Mostashari F, Mertz K, Hogin E, Atwal P. From the ONC: The strategy for advancing the exchange of health information	2012	tracking source of information (missing data), patient matching (privacy), workflow, liability
27	Steward W, Koester K, Collins A, Myers J. The essential role of reconfiguration capabilities in the implementation of HIV-related health information exchanges	2012	cost, technology gap, value, workflow
28	Deas TM, Solomon MR. Health information exchange: foundation for better care (Perspectives)	2012	cost, difficult to place value on HIE, missing data
29	Kralewski JE, Zink T, Boyle R. Factors Influencing Electronic Clinical Information Exchange in Small Medical Group Practices	2012	cost, lack of value, competition, technology gap, privacy

Art #	Study	Date	Barriers
30	Myers JJ, Koester KA, Chakravarty D, Pearson C, Maiorana A, Shade S, Steward W. Perceptions regarding the ease of use and usefulness of health information exchange systems among medical providers, case managers and nonclinical staff members working in HIV care and community settings	2012	usefulness (value), difficulty of interaction with HIE (tech support), workflow
31	Vest J, Jasperson JS. How are Health Professionals Using Health Information Exchange Systems? Measuring Usage for Evaluation and System Improvement	2012	effectiveness of Master Patient Index (MPI) (privacy), tech support
32	Adler-Milstein J, Bates DW, Jha AK. Operational Health Information Exchanges show substantial growth, but long-term funding remains a concern	2013	cost, provider pays while payer benefits, difficult to measure value
33	Dixon BE, Jones JF, Grannis SJ. Infection preventionists' awareness of and engagement in health information exchange to improve public health surveillance	2013	lack of awareness, decision support (workflow), usability, interoperability (standards), missing data
34	Furukawa MF, Patel V, Charles D, Swain M, Mostashari F. Hospital electronic health information exchange grew substantially in 2008-12	2013	limited interoperability (standards), competition, cost
35	Campion T, Edwards A, Johnson S, Kaushal R. Health information exchange system usage patterns in three communities: practice sites, users, patients, and data	2013	workflow
36	Miller A, Tucker C. Health information exchange, systems size and information silos	2013	standards, competition
37	Ben-Assuli O, Shabtia I, Leshno M. The impact of EHR and HIE on reducing avoidable admissions: controlling main differential diagnosis	2013	costs, missing data, decision making (workflow), leadership, competition
38	Vest JR, Campion TR, Kaushal R. Challenges, Alternatives, and Paths to Sustainability for Health Information Exchange Efforts	2013	cost, lack of value, competition
39	Thorn SA, Carter MA, Bailey JE. Emergency Physicians' Perspectives on Their Use of Health Information Exchange	2014	workflow, usability

Table 2. Barriers by the year published.

Barriers	2009	2010	2011	2012	2013	2014	Instances of the barrier	
Cost	4	2, 10, 20	12, 13, 22, 23	27, 28, 29	32, 34, 37, 38		15	15%
Efficiency/workflow		10, 20	13, 23, 24	14, 26, 27, 30	33, 35, 37	39	13	13%
Lack of technical support/tech gap		2, 10, 19	12, 13, 21, 23, 24, 25	27, 29, 30, 31			13	13%
Impedes competition		2, 10	12, 13	14, 28	34, 36, 37, 38		10	10%
Value of HIE is difficult to measure		18	21, 22, 25	27, 28, 29, 30	32, 38		10	10%
Privacy/security concerns		2, 18	23	14, 26, 29, 31			7	7%
Clinical data missing when needed			12, 24, 25	26, 28	33		6	6%
Usability		10	12, 13		33	39	5	5%
Heavily dependent on leadership of the organization		2	21, 22		37		4	4%
Liability concerns		2	23	14, 26			4	4%
Lack of standards					33, 34, 36		3	3%
Physician resistance	4	20					2	2%
Misaligned incentives			13		32		2	2%
Decreases quality		20					1	1%
Increases error		20					1	1%
Lack of awareness					33		1	1%
# barriers (n=16) 2, 13%	12, 75%	11, 69%	8, 50%	10, 63%	2, 13%	97		
# articles (n=28) 1, 4%	5, 18%	7, 25%	7, 25%	7, 25%	1, 4%			

In 2009, only 2 out of 16 barriers (13%) were listed by only 1 of 28 articles (4%). In 2010, 12 of 16 barriers (75%) were listed by 5 out of 28 articles (18%). In 2011, 11 of 16 barriers (69%) were listed by 7 of 28 articles (25%). In 2012, 8 of 16 barriers (50%) were listed by 7 of 28 articles (25%). In 2013, 7 out of 16 barriers (44%) were listed by 7 of 28 articles (25%). In 2014, 2 of 16 barriers (13%) were listed by only 1 of 28 articles (4%).

The barrier of cost was listed in the 2011 literature four times; one of which was a literature review. Cost was listed in the 2013 literature four times, the 2010 and 2012 literature three times, and in 2009 only once; in 2010-2012 one article was a literature review. If literature reviews were removed from the analysis, there would be a general increase in frequency of the barrier cost. From 2009-2014, the barrier "cost" was listed 15 of 97 instances (15%), "efficiency/workflow" and "lack of technology support / technology gap" were listed 13 of 97 instances (13%), "impedes competition," and "value of HIE is difficult to measure" were identified 10 of 97 instances (10%), "privacy/security issues" was listed 7 of 97 instances (7%), "clinical data missing" was listed 6 of 97 instances (6%), "usability" was listed 5 of 97 instances (5%), "heavily dependent on leadership" and "liability" were listed 4 of 97 instances (4%), "lack of standards" was listed 3 of 97 instances (3%), "physician resistance" and "misaligned incentives" were listed 2 of 97 instances (2%), "decreases quality," "increases error," and "lack of awareness" were each listed 1 of 97 instances (1%). The year 2010 revealed the greatest quantity of barriers listed (75%) and the years 2011-2013 tied for the number of articles published with barriers (25%).

Discussion

The concern of cost is discussed consistently in the literature, mostly with the concern of "no viable business plan" listed as the reason. This is not surprising. Very little participation in HIEs occurred prior to the ARRA in 2009, and most folded due to a lack of funding. The HITECH Act provided seed money, and the federal government asked the states to match or significantly contribute to the establishment of HIEs throughout the country. The stimulus money evaporates in 2014. By the end of 2014, HIEs will either develop a viable business plan or close their activities.

The second most consistent barriers discussed were efficiency/workflow, impedes competition, and value difficult to measure. These barriers were discussed four of the six years analyzed. The concept of participation in the HIE is intended to provide better quality care, but it makes no promises of efficiency. The concern that HIE participation will impede competition is concerning because it flies in the face of the altruistic nature of health care. This factor may be unique in the United States due to the competitive nature of the health care industry, the philosophy of health care as a privilege, and the

nonholistic definition of medicine that focuses on the identification and treatment of disease rather than promotion of overall health. Those health care organizations that treat Center for Medicaid and Medicare Services (CMS) beneficiaries submit for reimbursement based on diagnosis, which implies the presence of disease. The difficult nature of measuring the benefit of HIE is also not surprising. The use of HIE resources would not manifest itself in an obvious improvement of care; instead, the use of HIE would most likely result in fewer tests, for those that require tests, and the decrease of drug abuse for those who "shop" for controlled medications by frequent visitation of disparate emergency departments. The cost of participating in the HIE may indeed be more than the cost of duplicating the tests. In this regard, difficulty in measuring the value of the HIE may also point to the issue of cost.

The technical aspect of HIE was also listed frequently. This is an interesting item to be listed as a barrier which should have been addressed by the HITECH Act with the establishment of the Regional Extension Centers (RECs). These RECs provide technical support specifically to organizations transitioning to electronic health records and those interested in participating in an HIE. The frequency of this barrier dropped off after 2011, which could be a result of inter-organizational relationships being established with the operation of the RECs.

Privacy/security concerns is also not surprising. The Health Information Portability and Accountability Act (HIPAA) of 1996 creates an atmosphere of hypersensitivity for patient privacy and the security of health related information. An interesting observation is that the barrier "lack of standards" did not appear in the literature until 2013. This barrier could be accounted for by awareness, but the barrier "lack of awareness" also did not surface until the same year, although with only one third the frequency. Logically, an increase of awareness would result in an increase of standards development which should decrease the concern of privacy/security. The frequency of privacy/security did drop off after 2012, which might be indicative of the latter logic trail.

A limitation of this study is that it analyzes the frequency of barriers based on the year in which the articles were published, but it does not evaluate the year in which data were collected/analyzed. The editorial process varies by journal and quality of the initial draft. This could add 6 months or more to the publication process. A deeper analysis of the data-collection aspect of each article could make the focus of a future study. Another limitation is that the publication process will decay the internal validity of the study; for instance, more material will be published in 2014 during the editing/reviewing process that could contribute to the study.

The internal validity of this study seems otherwise sound. The inclusion of other literature reviews illustrates the exhaustive nature of the barriers mentioned. The only exception is the literature review by Joshi et al [20]. This article was published in 2010, but it could have easily focused on barriers mentioned prior to 2009, which was the earliest inclusion criteria for this study.

The external validity of this literature review seems strong. Studies included in this review included countries external to the United States; however, barriers in these countries might be juxtaposed to those in the United States due to the competitive nature of the health care industry in the United States.

The frequency of the barrier "cost" may identify problems in the future. By the end of 2014, federal funds for HIE initiatives will cease, which could cause the number of HIEs in the United States to plummet due to the lack of viable business plans. If the United States wants to ensure the longevity of HIEs in the country, it may need to lever additional incentives aimed at providers, or require health plans to contribute to the HIE programs. This raises several policy issues for the government to consider. Master patient indices at HIEs will only be effective if common data standards are in place across the nation.

The frequency of the barrier "efficiency/workflow" may be indicative of waste in the process of exchanging clinical data. Process improvement managers, or those familiar with Lean practices, need to map existing processes and take steps to eliminate wasted steps or procedures. By optimizing the process at the point of care, providers can feel confident that existing workflow procedures are efficient. If tools used to access data through HIEs are part of an inefficient process, the tools will simply transform them into expensive inefficient processes.

Additional measures could be taken by developers to alleviate the barrier "efficiency/workflow." Developers should increase their efforts to collect user needs and established workflows of users. Additional efforts at this step of software development could ensure ease of use for tools that access and contribute to HIE resources. These tools should be integrated into existing workflows, and the tools should be easily navigable. Accessing data through HIE should augment the effectiveness of care, and should not decrease the efficiency of care. The absence of concerns about privacy and security may only indicate the steady state of expectations of the same.

To address the barrier "lack of technical support," managers at RECs should focus closely on the local HIE efforts and reach out to the corresponding Regional Health Information Organizations (RHIOs). The RECs should help organizations realize improvements in both efficiency and effectiveness through the use of HIEs. The managers at Regional Health Information Organizations (RHIOs) should realize that the services that they provide should not take any longer to access than the repeated tests that the HIEs are supposed to mitigate. Managers at RHIOs should also reach out to senior leadership at organizations that could participate in HIEs to win their confidence. Once senior leadership is convinced of the value of HIE, our nations should see additional participation and inter-organizational trust that would overcome competitive environments.

Acknowledgments

We would like to thank the reviewers for making this a better manuscript. Thank you for your constructive feedback and input that was both educational and professional.

Authors' Contributions

This manuscript was the result of directed research and collaborative effort between graduate students and a professor in the Masters of Health Administration program.

Conflicts of Interest

None declared.

References

1. Health Information Technology for Economic Clinical Health Act, 42 U. C. S. URL: http://www.healthit.gov/sites/default/files/hitech_act_excerpt_from_arra_with_index.pdf [accessed 2014-09-10] [WebCite Cache ID 6SUhqPDHe]
2. Fontaine P, Ross SE, Zink T, Schilling LM. Systematic review of health information exchange in primary care practices. J Am Board Fam Med 2010;23(5):655-670 [FREE Full text] [doi: 10.3122/jabfm.2010.05.090192] [Medline: 20823361]
3. Walker J, Pan E, Johnston D, Adler-Milstein J, Bates DW, Middleton B. The value of health care information exchange and interoperability. Health Aff (Millwood) 2005;Suppl Web Exclusives:W5-10 [FREE Full text] [doi: 10.1377/hlthaff.w5.10] [Medline: 15659453]
4. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, et al. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health Aff (Millwood) 2005 Oct;24(5):1103-1117 [FREE Full text] [doi: 10.1377/hlthaff.24.5.1103] [Medline: 16162551]
5. Frisse ME, Holmes RL. Estimated financial savings associated with health information exchange and ambulatory care referral. J Biomed Inform 2007 Dec;40(6 Suppl):S27-S32 [FREE Full text] [doi: 10.1016/j.jbi.2007.08.004] [Medline: 17942374]
6. Adler-Milstein J, McAfee AP, Bates DW, Jha AK. The state of regional health information organizations: current activities and financing. Health Aff (Millwood) 2008 Feb;27(1):w60-w69 [FREE Full text] [doi: 10.1377/hlthaff.27.1.w60] [Medline: 18073225]
7. Vest JR, Gamm LD. Health information exchange: persistent challenges and new strategies. J Am Med Inform Assoc 2010 Jun;17(3):288-294 [FREE Full text] [doi: 10.1136/jamia.2010.003673] [Medline: 20442146]
8. Kho AN, Lemmon L, Commiskey M, Wilson SJ, McDonald CJ. Use of a regional health information exchange to detect crossover of patients with MRSA between urban hospitals. J Am Med Inform Assoc 2008 Apr;15(2):212-216 [FREE Full text] [doi: 10.1197/jamia.M2577] [Medline: 18096903]
9. Committee on Quality of Health Care in America, Institute of Medicine. Crossing the quality chasm: a new health system for the 21st century. Washington, D.C: National Academy Press; 2001.
10. Ross SE, Schilling LM, Fernald DH, Davidson AJ, West DR. Health information exchange in small-to-medium sized family medicine practices: motivators, barriers, and potential facilitators of adoption. Int J Med Inform 2010 Feb;79(2):123-129. [doi: 10.1016/j.ijmedinf.2009.12.001] [Medline: 20061182]
11. Tham E, Ross SE, Mellis BK, Beaty BL, Schilling LM, Davidson AJ. Interest in health information exchange in ambulatory care: a statewide survey. Appl Clin Inform 2010;1(1):1-10 [FREE Full text] [doi: 10.4338/ACI-2009-10-RA-0007] [Medline: 23616824]
12. Rudin R, Volk L, Simon S, Bates D. What Affects Clinicians' Usage of Health Information Exchange? Appl Clin Inform 2011 Jan 1;2(3):250-262 [FREE Full text] [doi: 10.4338/ACI-2011-03-RA-0021] [Medline: 22180762]
13. Patel V, Abramson EL, Edwards A, Malhotra S, Kaushal R. Physicians' potential use and preferences related to health information exchange. Int J Med Inform 2011 Mar;80(3):171-180. [doi: 10.1016/j.ijmedinf.2010.11.008] [Medline: 21156351]
14. Pevnick JM, Claver M, Dobalian A, Asch SM, Stutman HR, Tomines A, et al. Provider stakeholders' perceived benefit from a nascent health information exchange: a qualitative analysis. J Med Syst 2012 Apr;36(2):601-613 [FREE Full text] [doi: 10.1007/s10916-010-9524-x] [Medline: 20703673]
15. URL: http://dashboard.healthit.gov/hitadoption/ [accessed 2014-06-10] [WebCite Cache ID 6QEk8CEMk]
16. eHealth Initiative. Results from survey on health data exchange 2013 URL: http://www.ehidc.org/resource-center/surveys/view_document/333-survey-results-results-from-survey-on-data-exchange-2013-data-exchange [accessed 2014-06-10] [WebCite Cache ID 6QElBOTki]
17. U.S. Health and Human Services. Data exchange growing through EHR adoption, new study finds URL: http://www.hhs.gov/news/press/2013pres/08/20130805a.html [accessed 2014-06-10] [WebCite Cache ID 6QElkTR9E]
18. Wright A, Soran C, Jenter CA, Volk LA, Bates DW, Simon SR. Physician attitudes toward health information exchange: results of a statewide survey. J Am Med Inform Assoc 2010 Feb;17(1):66-70 [FREE Full text] [doi: 10.1197/jamia.M3241] [Medline: 20064804]

19. Dixon BE, Zafar A, Overhage JM. A Framework for evaluating the costs, effort, and value of nationwide health information exchange. J Am Med Inform Assoc 2010 Jun;17(3):295-301 [FREE Full text] [doi: 10.1136/jamia.2009.000570] [Medline: 20442147]
20. Clinical Value-Add for Health Information Exchange (HIE). IJMI 2011 Jan;6(1). [doi: 10.5580/1b89]
21. Korst LM, Aydin CE, Signer JM, Fink A. Hospital readiness for health information exchange: development of metrics associated with successful collaboration for quality improvement. Int J Med Inform 2011 Aug;80(8):e178-e188 [FREE Full text] [doi: 10.1016/j.ijmedinf.2011.01.010] [Medline: 21330191]
22. Adler-Milstein J, Bates DW, Jha AK. A survey of health information exchange organizations in the United States: implications for meaningful use. Ann Intern Med 2011 May 17;154(10):666-671. [doi: 10.7326/0003-4819-154-10-201105170-00006] [Medline: 21576534]
23. Lluch M. Healthcare professionals' organisational barriers to health information technologies-a literature review. Int J Med Inform 2011 Dec;80(12):849-862. [doi: 10.1016/j.ijmedinf.2011.09.005] [Medline: 22000677]
24. Gadd CS, Ho YX, Cala CM, Blakemore D, Chen Q, Frisse ME, et al. User perspectives on the usability of a regional health information exchange. J Am Med Inform Assoc 2011 Oct;18(5):711-716 [FREE Full text] [doi: 10.1136/amiajnl-2011-000281] [Medline: 21622933]
25. Hincapie AL, Warholak TL, Murcko AC, Slack M, Malone DC. Physicians' opinions of a health information exchange. J Am Med Inform Assoc 2011 Feb;18(1):60-65 [FREE Full text] [doi: 10.1136/jamia.2010.006502] [Medline: 21106994]
26. Williams C, Mostashari F, Mertz K, Hogin E, Atwal P. From the Office of the National Coordinator: the strategy for advancing the exchange of health information. Health Aff (Millwood) 2012 Mar;31(3):527-536. [doi: 10.1377/hlthaff.2011.1314] [Medline: 22392663]
27. Steward WT, Koester KA, Collins SP, Maiorana A, Myers JJ. The essential role of reconfiguration capabilities in the implementation of HIV-related health information exchanges. Int J Med Inform 2012 Oct;81(10):e10-e20. [doi: 10.1016/j.ijmedinf.2012.07.004] [Medline: 22841703]
28. Deas TM, Solomon MR. Health information exchange: foundation for better care. Gastrointest Endosc 2012 Jul;76(1):163-168. [doi: 10.1016/j.gie.2012.03.1406] [Medline: 22726476]
29. Kralewski JE, Zink T, Boyle R. Factors influencing electronic clinical information exchange in small medical group practices. J Rural Health 2012 Jan;28(1):28-33. [doi: 10.1111/j.1748-0361.2011.00372.x] [Medline: 22236312]
30. Myers JJ, Koester KA, Chakravarty D, Pearson C, Maiorana A, Shade SB, et al. Perceptions regarding the ease of use and usefulness of health information exchange systems among medical providers, case managers and non-clinical staff members working in HIV care and community settings. Int J Med Inform 2012 Oct;81(10):e21-e29. [doi: 10.1016/j.ijmedinf.2012.07.005] [Medline: 22854159]
31. Vest JR, Jasperson '. How are health professionals using health information exchange systems? Measuring usage for evaluation and system improvement. J Med Syst 2012 Oct;36(5):3195-3204 [FREE Full text] [doi: 10.1007/s10916-011-9810-2] [Medline: 22127521]
32. Adler-Milstein J, Bates DW, Jha AK. Operational health information exchanges show substantial growth, but long-term funding remains a concern. Health Aff (Millwood) 2013 Aug;32(8):1486-1492. [doi: 10.1377/hlthaff.2013.0124] [Medline: 23840051]
33. Dixon BE, Jones JF, Grannis SJ. Infection preventionists' awareness of and engagement in health information exchange to improve public health surveillance. Am J Infect Control 2013 Sep;41(9):787-792. [doi: 10.1016/j.ajic.2012.10.022] [Medline: 23415767]
34. Furukawa MF, Patel V, Charles D, Swain M, Mostashari F. Hospital electronic health information exchange grew substantially in 2008-12. Health Aff (Millwood) 2013 Aug;32(8):1346-1354. [doi: 10.1377/hlthaff.2013.0010] [Medline: 23918477]
35. Campion TR, Edwards AM, Johnson SB, Kaushal R, HITEC investigators. Health information exchange system usage patterns in three communities: practice sites, users, patients, and data. Int J Med Inform 2013 Sep;82(9):810-820. [doi: 10.1016/j.ijmedinf.2013.05.001] [Medline: 23743323]
36. Miller AR, Tucker C. Health information exchange, system size and information silos. J Health Econ 2014 Jan;33:28-42. [doi: 10.1016/j.jhealeco.2013.10.004] [Medline: 24246484]
37. Ben-Assuli O, Shabtai I, Leshno M. The impact of EHR and HIE on reducing avoidable admissions: controlling main differential diagnoses. BMC Med Inform Decis Mak 2013;13:49 [FREE Full text] [doi: 10.1186/1472-6947-13-49] [Medline: 23594488]
38. Vest JR, Campion TR, Kaushal R, HITEC Investigators. Challenges, alternatives, and paths to sustainability for health information exchange efforts. J Med Syst 2013 Dec;37(6):9987. [doi: 10.1007/s10916-013-9987-7] [Medline: 24141531]
39. Thorn SA, Carter MA, Bailey JE. Emergency physicians' perspectives on their use of health information exchange. Ann Emerg Med 2014 Mar;63(3):329-337. [doi: 10.1016/j.annemergmed.2013.09.024] [Medline: 24161840]

Abbreviations

ARRA: American recovery and reinvestment act
CFR: code of federal regulation

CINAHL: cumulative index for nursing and allied health literature
CMS: Center for Medicaid and Medicare services
EHR: electronic health record
EMR: electronic medical record
HC: health care
HIE: health information exchange
HIPAA: health insurance portability and accountability act
HIT: health information technology
HITECH: health information technology for economic and clinical health
MPI: master patient index
IOM: institute of medicine
IRB: institutional review board
MRSA: Methicillin-resistant Staphylococcus aureus
ONC: office of the national coordinator
REC: regional extension center
RHIO: regional health information organization

Edited by G Eysenbach; submitted 19.06.14; peer-reviewed by L Fulton, R Elmore; comments to author 14.08.14; revised version received 15.08.14; accepted 01.09.14; published 30.09.14.

Please cite as:
Kruse CS, Regier V, Rheinboldt KT
Barriers Over Time to Full Implementation of Health Information Exchange in the United States
JMIR Med Inform 2014;2(2):e26
URL: http://medinform.jmir.org/2014/2/e26/
doi:10.2196/medinform.3625
PMID:25600635

Original Paper

Health Information Exchange Implementation: Lessons Learned and Critical Success Factors From a Case Study

Sue S Feldman[1,2,3*], RN, MEd, PhD; Benjamin L Schooley[2,4*], MBA, PhD; Grishma P Bhavsar[2*], MPH

[1]Central Virginia Health Network, Richmond, VA, United States

[2]Arnold School of Public Health, University of South Carolina, Columbia, SC, United States

[3]School of Information Systems and Technology, Claremont Graduate University, Claremont, CA, United States

[4]Department of Integrated Information Technology, University of South Carolina, Columbia, SC, United States

[*]all authors contributed equally

Corresponding Author:
Sue S Feldman, RN, MEd, PhD
Central Virginia Health Network
4900 Cox Road #200
Richmond, VA, 23060
United States
Phone: 1 6616188805
Fax: 1 6616188805
Email: suefeldman1009@gmail.com

Abstract

Background: Much attention has been given to the proposition that the exchange of health information as an act, and health information exchange (HIE), as an entity, are critical components of a framework for health care change, yet little has been studied to understand the value proposition of implementing HIE with a statewide HIE. Such an organization facilitates the exchange of health information across disparate systems, thus following patients as they move across different care settings and encounters, whether or not they share an organizational affiliation. A sociotechnical systems approach and an interorganizational systems framework were used to examine implementation of a health system electronic medical record (EMR) system onto a statewide HIE, under a cooperative agreement with the Office of the National Coordinator for Health Information Technology, and its collaborating organizations.

Objective: The objective of the study was to focus on the implementation of a health system onto a statewide HIE; provide insight into the technical, organizational, and governance aspects of a large private health system and the Virginia statewide HIE (organizations with the shared goal of exchanging health information); and to understand the organizational motivations and value propositions apparent during HIE implementation.

Methods: We used a formative evaluation methodology to investigate the first implementation of a health system onto the statewide HIE. Qualitative methods (direct observation, 36 hours), informal information gathering, semistructured interviews (N=12), and document analysis were used to gather data between August 12, 2012 and June 24, 2013. Derived from sociotechnical concepts, a Blended Value Collaboration Enactment Framework guided the data gathering and analysis to understand organizational stakeholders' perspectives across technical, organizational, and governance dimensions.

Results: Several challenges, successes, and lessons learned during the implementation of a health system to the statewide HIE were found. The most significant perceived success was accomplishing the implementation, although many interviewees also underscored the value of a project champion with decision-making power. In terms of lessons learned, social reasons were found to be very significant motivators for early implementation, frequently outweighing economic motivations. It was clear that understanding the guides early in the project would have mitigated some of the challenges that emerged, and early communication with the electronic health record vendor so that they have a solid understanding of the undertaking was critical. An HIE implementations evaluation framework was found to be useful for assessing challenges, motivations, value propositions for participating, and success factors to consider for future implementations.

Conclusions: This case study illuminates five critical success factors for implementation of a health system onto a statewide HIE. This study also reveals that organizations have varied motivations and value proposition perceptions for engaging in the exchange of health information, few of which, at the early stages, are economically driven.

(JMIR Med Inform 2014;2(2):e19) doi:10.2196/medinform.3455

KEYWORDS

health information exchange; blended value collaboration enactment framework; HIE value proposition; interorganizational systems; HIE implementation

Introduction

Investing in Health Information Technology

The Health Information Technology for Economic and Clinical Health, HITECH, Act, enacted as part of the American Recovery and Reinvestment Act of 2009, set a national goal of investing in health information technology to improve health care delivery. To meet this goal, the electronic exchange of health information between providers is essential to ensure coordinated, efficient, and quality care. This exchange can be accomplished through various local, regional, or statewide organizations that build the infrastructure to facilitate secure health information exchange (HIE); the federal government has already entered into cooperative agreements with 56 states and territories to fund infrastructure to enable these efforts [1]. HIE is a complex and emergent system of structures and actions, which varies in scope and scale. The current study discusses HIE as both the *act* of exchanging health information between two collaborating organizations, and the *entity* that facilitates such exchange. The goal of this study, which focuses on the implementation of HIE as an interorganizational health care system, is to understand the organizational motivations and value propositions apparent during HIE implementation. These value propositions are analyzed through an interorganizational system (IOS) information technology (IT) governance lens that considers the technical, organizational, and governance dimensions of HIE value. We apply the framework to: (1) evaluate HIE implementation challenges, successes, and lessons learned; and (2) extract value propositions across organizational stakeholders.

Health Information Exchange

Much attention has been given to the proposition that the exchange of health information is a critical component of a framework for health care change, the Triple Aim being: (1) better patient experiences through quality and satisfaction; (2) better health outcomes of populations; and (3) reduction of per capita cost of health care [2]. These changes will rely on organizational entities that have entered cooperative agreements with the federal government to provide technical infrastructure, organizational structure, and governance mechanisms for completing the act of HIE [1]. The act of HIE, described various ways in the literature, can be conducted across affiliated physicians' offices, hospitals, and clinics; or can occur between completely disparate systems [3,4]. HIE across disparate systems allows clinical information to follow patients as they move across different care settings, whether or not they share an organizational affiliation. For example, this might include a hospital or health system connected to an HIE network that is, in turn, connected to several for-profit and not-for-profit competing hospitals or health systems networks. On a broad scale, this type of HIE holds great promise for achieving the Triple Aim goals.

Health Information Exchange Benefits and Challenges

The implementation and use of HIE technology have influenced patient care by allowing providers direct access to health information, reducing time to obtain health information, and increasing providers' awareness of patient interactions with the health care system [5]. The benefits and challenges of HIE have been studied in prior research. Regarding patient experiences, previous studies have found improved coordination of care and enhanced patient health outcomes for human immunodeficiency virus patients [6], higher patient satisfaction [7], informed patient care [8], efficient care [8], and positive patient perception of impact on care coordination [9]. However, others have found that the benefits of HIE in relation to patient outcomes are limited [10].

From a broader provider and patient perspective, timely sharing of a patient's clinical information can improve the accuracy of diagnoses, reduce the number of duplicative tests, prevent hospital readmissions, and prevent medication errors [3,11,12]. From a public health perspective, the exchange of health information has fostered positive relationships with public health agencies [13], improved public health surveillance [14], and increased the efficiency and quality of public health reporting [15].

Though the theoretical case for HIE on reducing the utilization and cost of health care services is compelling and has received a great deal of emphasis [16], empirical evidence is still inconclusive [17,18]. This may reflect the nascent nature of HIE, especially between disparate systems, and the fact that these systems and the context are complex and emergent. For example, HIE has faced challenges like those of other new IT initiatives, disparate and noninteroperable information technologies [19,20]; a range of technical, work flow, and organizational challenges to exchanging information [17,21]; and a variety of governance challenges [22,23]. Yet the HIEs that have continued to operate have done so with evolving and maturing technical, organizational, and governance structures [24-27].

Still, much more research is needed to understand HIE, how it operates, what factors contribute to success, and even how success should be defined. At this early phase of HIE development and implementation, it is important to study the system and its context to improve upon existing methods, tools, and frameworks. This study investigates the value of HIE from an IT implementation perspective. Specifically, it asks what motivations, challenges, and successes lead to value realization across organizations working together to on-board to a state HIE? The sociotechnical systems (STS) approach of this study applies an IOS framework developed through previous work to understand blended value across participating organizations.

Sociotechnical Systems Approach

An STS approach examines social/community links to the technical [28]. STS design includes several levels of abstraction including mechanical (hardware), informational (software), psychological (person), and social (community). Such an inclusive approach is aimed at understanding interdependent linkages between increasingly complex social and technological components. Working together, these components consider social motivations and accomplish a set of social goals that otherwise would not be realized.

The highest-order social benefit (human life) of health information sharing are stated quite succinctly by Porter and Teisburg [29], "The social benefits of results information will be even greater in health care than in the financial markets, because the physical well being of Americans is at stake." Social value factors include the range of intangible and actor-based or organizational considerations that contribute to collaboration success. Furthermore, two studies almost 20 years apart suggest that successful IT project advancement is frequently associated with social elements [30,31]. In his book on infrastructure delivery in public-private collaborations, Mody [32] draws from the railway and transportation systems examples to suggest that social considerations, such as being able to deliver goods and information to the right place at the right time, might exceed those of economic returns and could exert greater significant pressure.

Therefore, examination of the social motivations and benefits deserve to be considered in a different light than a customary return on investment model commonly considered in information exchange in the business world. This blended value proposition has been defined as the combination of social and economic value used to maximize total returns where "the core nature of investment and return is not a tradeoff between social and financial interest, but rather the pursuit of an embedded value proposition composed of both" [33]. Emerson [33] continues,

> *Societies cannot function strictly on the basis of their economic enterprise. It is social commerce that allows individuals and institutions to pursue the traditional financial returns sought by mainstream financial capital market players.* [J Emerson]

An STS approach, which focuses on systems that are both technologically sound and socially sustainable [34], has been applied to the study of HIE because of the multiple organizations, user types, hardware and software technologies, and sociopolitical motivations and goals involved in its composition. Though relatively few studies have examined an HIE network in operation, a sociotechnical approach was previously applied and shown to be appropriate to the study of HIE [5].

Interorganizational Systems Implementations

An IOS is an IT-based system shared by two or more independent organizations [35]. Prior research on IOSs focused on the cross-organizational features of an STS [35,36]. While the implementation of IOS has been studied for decades across a wide array of industries, few studies have addressed health care, and fewer still have addressed HIE. IOS studies show the importance of: (1) learning from early adopters [37], and (2) evaluating the process of implementation to understand lessons learned and the real and perceived value of an IOS [38,39].

From Interorganizational System to Health Information Exchange Implementations

Evaluations of information system collaboration require looking beyond a single focus and attending to multiple dimensions [40]. This perspective acknowledges that the collaboration of multiple stakeholders may hold the potential to create something new and better, as well as to create public value [41]. Similarly, a multidimensional perspective is required in evaluating the exchange of health information [42].

Evaluations of HIE and the benefits and challenges of exchanging health information have been studied in various contexts. For example, a 2011 study suggested that US $2 million in uncompensated care cost recovery is achievable with use of the nationwide HIE (now eHealth Exchange) as applied to disability determination [43]; and a more recent study estimated the resource utilization impacts resulting from using eHealth Exchange for emergency department visits [44]. Yet, few studies exist regarding the value of HIE at a statewide level. While these economic findings are important and may drive a sustainable IOS, understanding social value or motivation is important to HIE implementation.

As states end their cooperative agreements with the federal government, it is helpful to understand the challenges, successes, and lessons learned from an early on-boarder or implementation. While literature exists about information systems implementations across various industries, little is known about health systems implementations between public-private entities (eg, a private hospital and state HIE).

The aim of this case study is to provide insight into the technical, organizational, and governance aspects of a large private health system (Inova Health System) and the Virginia statewide HIE (ConnectVirginia EXCHANGE), organizations with the shared goal of exchanging health information. In this case study, the Blended Value Collaboration Enactment Framework, a multidimensional value framework [45], discussed later in this paper, provided a conceptual framework for evaluating the implementation process by which an organization becomes connected to a system to facilitate the exchange of information (ie, on-boarding), the first, to our knowledge, on-boarding to ConnectVirginia EXCHANGE.

Methods

Overview

The study design comprised direct observation, informal information gathering, document analysis, and semistructured interviews to study HIE implementation across technical, organizational, and governance dimensions. The study assessed the first, to our knowledge, on-boarding of a health care system onto the Virginia statewide HIE, ConnectVirginia EXCHANGE, using a formative evaluation of the implementation phase of the systems development life cycle. The study did not address

the exchange of information, but rather the process of HIE implementation.

Study Setting and Background

ConnectVirginia EXCHANGE

In March 2010, the Office of the National Coordinator for Health IT (ONC) awarded state cooperative agreements to the states and territories in the United States to develop infrastructure supporting the electronic exchange of health information. At that time, the Virginia Department of Health (VDH), the state-designated entity for Virginia, was awarded US $11.6 million. In September 2011, Community Health Alliance (CHA) was awarded a contract from VDH to build the Virginia Statewide HIE, ConnectVirginia. The organization to accomplish this statewide was subsequently initiated. Statewide HIEs were required to enable information exchange using standardized technologies, tools, and methods. Between September 2011 and February 2014, ConnectVirginia designed, tested, developed, and implemented three technical exchange services: (1) ConnectVirginia DIRECT Messaging (a secure messaging system), (2) ConnectVirginia EXCHANGE (the focus of this study), and (3) a Public Health Reporting Pathway (a pathway with VDH for public health reporting). ConnectVirginia EXCHANGE is a query/retrieve service in which a deliberate query passively returns one or more standardized continuity of care documents (CCDs; these provide a means of sharing standardized health data between organizations) from any other "system" on-boarded and connected to ConnectVirginia. This study examines and reports on the first implementation to ConnectVirginia EXCHANGE by Inova Health System (Inova).

Inova Health System

Inova [46] primarily serves the Northern Virginia and Washington, DC, markets and includes five hospitals with more than 1700 licensed beds and 16,000 employees. This comprehensive network of inpatient hospitals also includes outpatient services and facilities, primary and specialty care physician practices, and health and wellness initiatives. The inpatient facilities use a well known electronic health record (EHR), and the affiliated outpatient practices have access to that EHR. In keeping with Inova's vision to increase value for patients and build an integrated network within and outside of their own hospitals, Inova became the first node to on-board to the ConnectVirginia EXCHANGE.

Electronic Health Record System

The EHR software involved in this study is primarily for mid-size and large medical groups, hospitals, and integrated health care organizations spanning clinical, access, and revenue functions. It provides an intranetwork data-sharing pathway (this specific EHR to this specific EHR), as well as an external data-sharing pathway (this specific EHR to a different EHR or system). Use of the external data-sharing pathway is the subject of this implementation study.

Research Process

ConnectVirginia initiated and managed the on-boarding process. The on-boarding process for Inova began with a kick-off meeting on August 2, 2012, and concluded with a test to exchange electronic documents with ConnectVirginia on April 26, 2013 (184 total workdays). Along with Inova and ConnectVirginia, implementation involved two critical subcontractors: (1) MEDfx (the software vendor for ConnectVirginia), and (2) MedVirginia (the CCD content consultant). Figure 1 shows the evaluation timeline.

Figure 1. Evaluation timeline.

Formative Evaluation of Information Systems Implementations

To assess the HIE implementation process from the perspective of an STS, this study used a formative evaluation methodology. Formative evaluations are widely used in young and developing initiatives to enable continuous improvement throughout the development and implementation stages [47,48]. From a practical perspective, this approach allows organizations to learn from past mistakes and develop better methods for assessing success [42,48]. This methodology allowed researchers to investigate the first implementation of ConnectVirginia EXCHANGE for a new and emergent type of system (ie, HIE) that is rapidly expanding across thousands of US health care systems.

To study IT implementations, Cooper and Zmud [49] proposed a diffusion process model of IT implementation that includes factors influencing implementation. Their model captures both the process and its context, subcategorized into stages. For example, their diffusion process model of IT implementation proposes six stages: (1) initiation, (2) adoption, (3) adaptation, (4) acceptance, (5) routinization, and (6) infusion. The present HIE evaluation covers only the *adoption* and *adaptation* stages of Cooper and Zmud's model, which include: (1) gaining organizational backing for implementing IT applications, and (2) developing and installing IT applications, and developing and revising organizational policies and procedures for ongoing use of the IT applications. The classic system development life cycle recognizes four distinct implementation phases that can be used in IT evaluations: (1) preimplementation, (2) during

implementation, (3) postimplementation, and (4) routine operation [42]. Evaluations such as this study conducted on the "during implementation" stage, use qualitative methodology [47,48]. It has been suggested that evaluations during this stage may be more important than those providing proof of outcomes [50], as the former can provide guidelines and lessons learned for others. The methods applied herein aim to extract valuable measures, and disseminate lessons learned for other HIE implementation efforts.

Information Technology Implementation Measures

Evaluations of the exchange of health information can be challenging [51], partly due to the lack of any single model for HIE [50]. Implementation measures are generally chosen for their value to stakeholders [52]. Evaluations should determine not only how well a system works, but also how well it works for particular users in a particular setting [42].

Several measures have been applied to evaluations of IT implementations. Categories span different levels of abstraction including: (1) technical, (2) organizational, and (3) governance. In prior research, implementation measures pertaining to both technical and organizational dimensions included: (1) degree and type of data usage [50,53-55], (2) level of complexity of business processes [56], (3) completeness of information [47,50,54], (4) resistance to change [56], (5) unintended consequences [50,53], and (6) facilitators [47] and barriers [47,50] to implementation. Organizational and governance dimensions in implementations include: (1) communication [47], (2) trust [47], (3) organizational structure [53], (4) sustainability [12,54,57,58], (5) roles and power relations between participants [56], (6) levels of leadership commitment [47], and (7) representativeness and motivations of stakeholders [47,50,57].

The technical, organizational, and governance aspects of HIE, as well as their interactions with each other; provide a basis for the evaluation measurements currently utilized [42]. These measures were applied to the study of ConnectVirginia EXCHANGE within the context of a previously tested analytical framework for HIE [43].

Analytical Framework

Enactment theory describes how people act within organizations [59]. When people carry out an act, they take into account their past experiences, events, and structures; determine a course of action; and then set that course into action. It is a form of social construction. Fountain's technology enactment framework builds on enactment theory and considers that technical factors and organizational structures are embedded *within* each collaborating organization, and that the relationship between multiple factors is critical [60]. Others have suggested that while technical performance is a crucial element in any resulting information exchange between organizations, successful interorganizational data exchanges frequently hinge on organizational and governance factors [56,61-63]. However, other research notes that motivational factors and context can be the true underpinnings of collaboration [64,65]. As such, collaborations for information exchange require organizations to look beyond a single focus and give attention to multiple dimensions of collaboration [40].

Based on the aforementioned work of Fountain, Schooley, and Emerson, and because of its prior use and demonstrated utility in assessing multi-organizational HIE efforts, the Blended Value Collaboration Enactment Framework was used to guide implementation and evaluation [45] (Figure 2 shows this framework). Framework development drew upon STS concepts and frameworks. Its importance for this study is that the framework: (1) considers each organizational stakeholder and its respective social and economic motivations for participating in HIE; (2) differentiates between technical, organizational, and governance dimensions; and (3) focuses on determining value propositions across stakeholders. For this study, and within the context of HIE, technical is defined as elements associated with the system or infrastructure; organizational is defined as elements associated with any and all of the stakeholders; and governance is defined as elements associated with decision making [45].

The above framework also considers the *value proposition* of HIE across stakeholders, including the social and economic motivations that lead to a more successful and sustainable HIE. A value proposition can be defined as the implicit promise of mutual value to the organization and its customers and/or partners [66]. For example, an in-depth case study of the fashion industry found that interorganizational value propositions could have both "hard" elements (economic gain, technological mastery, etc) and "soft" elements (brand identity, trust relationships, etc) [67]. Past research on HIE has illustrated that each stakeholder organization has its own value-driven motivations for participating in the exchange of health information, social including clinical (eg, "Is this the right thing to do for public health and wellness?") versus economic (eg, "How does this impact our financial bottom line?").

The intended output of the Blended Value Collaboration Enactment Framework is the resulting system performance [45]. Since this study reports on the implementation of HIE and not on its actual use (from which system performance would be derived), the "output" section of Figure 2 has been modified in Figure 3 to reflect critical success factors, as is more appropriate for implementation studies [68]. The framework also proposes that motivations and value propositions may change of over time (T_1 and T_2 in Figure 2). This study investigates only the implementation stage and does not evaluate how these dimensions change over time. Therefore, only the unshaded areas of the framework are germane to this analysis.

Figure 2. Blended Value Collaboration Enactment Framework [45]. T1 and T2= changes over time.

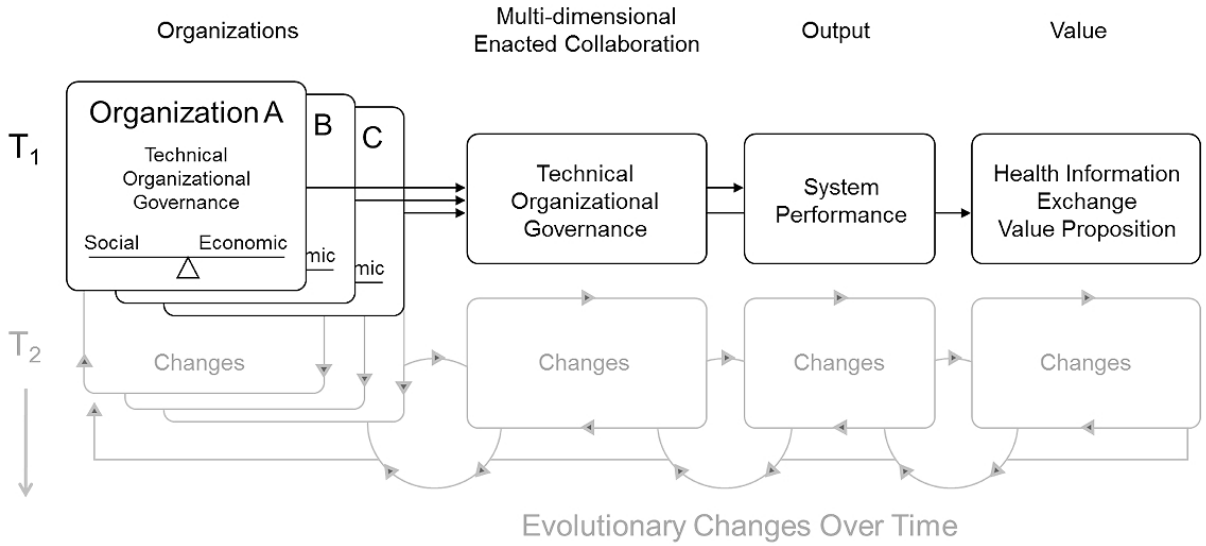
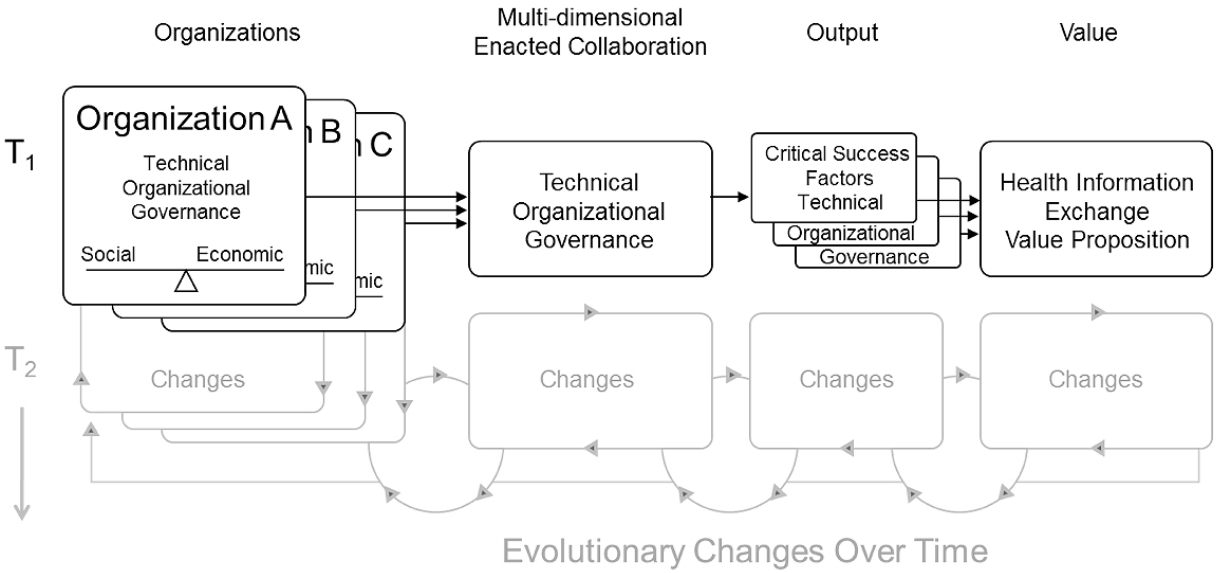


Figure 3. Blended Value Collaboration Enactment Framework. T1 and T2=changes over time.



Observation, Informal Information Gathering, and Document Analysis

Data collection took place August 2012-June 2013 by one of the authors, who was the external evaluator to ConnectVirginia and not part of the implementation team. A total of 36 hours of observation of the on-boarding process occurred across planning, coordination, implementation, and problem-solving meetings held either in-person or via conference calls. Each organization was represented in each meeting, and the meetings provided an environment for conducting informal information gathering. Various documents, such as meeting notes and detailed meeting minutes, were also analyzed.

Semistructured Interviews

Qualitative methods were employed to understand *how* and *why* the factors in each dimension contribute to or influence the overall implementation and value derived. At the end of the project (between May 8 and June 24, 2013), 12 60-minute, semistructured in-person interviews were conducted across the five participating organizations (ConnectVirginia, MEDfx, MedVirginia, EHR vendor, and Inova). Table 1 provides additional details about the interviewees. Purposive sampling was used to select interviewees based on: (1) their belonging to one of the above-mentioned organizations, and (2) their degree of participation in the on-boarding process. For example, individuals who participated in the majority of project manager (PM), technical, or system testing meetings were invited for interviews, and all invitees agreed to participate. Persons such as consumers of the exchanged health information (ie, clinicians) were not interviewed because, at the time of the study, there was no routine exchange of information for real-world use. Interviews were conducted by one of the authors with expertise in conducting interviews, and who was also the external evaluator to ConnectVirginia and not part of the implementation team.

Interview questions were designed to develop a clearer picture of the on-boarding process, to recreate the actual timeline, and

to support information from calls, documents, and informal information gathering. Table 2 provides a sampling of interview questions. Not all questions were appropriate for all interviewees, and therefore were not asked to all interviewees. Likewise, if interviewees had in-depth knowledge of a particular process, secondary questions were asked that may or may not have been used for other interviewees.

Table 1. Interviewee's by organization, position, and role.

Organization	Position	Role during implementation
ConnectVirginia	Executive Director	Oversight
	PM	Daily operations management of the implementation
MEDfx	Chief Operations Officer	Oversight
	PM	Daily operations management of the implementation
MedVirginia	Chief Information Officer	Oversight
	Systems Analyst	CCD content validation
EHR vendor	Application Support Specialist (x3)	Provided vendor support during the implementation
Inova	Executive Vice President and Chief Technology Officer	Oversight and internal champion
	Senior Vice President and Chief Information Officer	Oversight
	PM	Daily operations management of the implementation

Table 2. Sample interview questions.

Sample interview question types	
Technical	
	What were the initial technical processes involved in on-boarding to ConnectVirginia?
	What technical advances and/or information could have streamlined the on-boarding process?
	What technical challenges emerged and how were they addressed?
	Were any technical "workarounds" employed? If so, please explain.
	What technical processes were particularly successful and why?
	To what extent was the technical assistance that you received helpful?
	Please describe your current level of HIE (eg, within your organization, outside your organization, labs, etc).
Organizational	
	To what extent did organizational leadership impact the on-boarding process?
	What is the value proposition of on-boarding to ConnectVirginia?
	What organizational challenges emerged and how were they addressed?
	What is needed to have HIE become a standard of care?
	What was particularly successful regarding organizational leadership?
Governance	
	What were the key elements of the governance structure within your organization for on-boarding to ConnectVirginia?
	What governance structures do you see as vital for sustainability or growth of HIE across ConnectVirginia?
	To what extent were on-boarding guides governing implementation useful, helpful, or challenging?

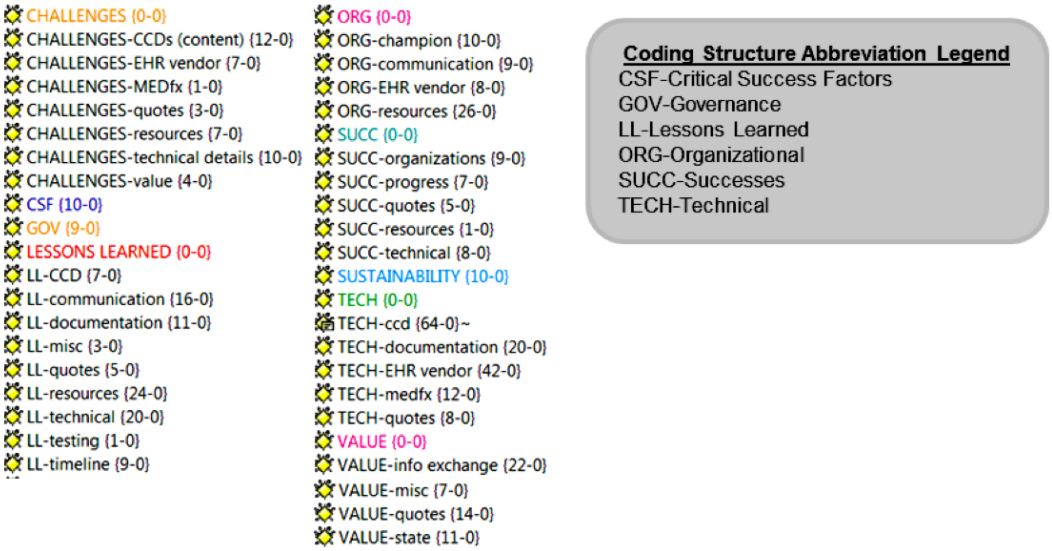
Data Analysis

Interviews were transcribed verbatim and imported into ATLAS.ti, a qualitative data analysis software application [69]. The Blended Value Collaboration Enactment Framework was used to guide the analysis. Each dimension from the framework (technical, organizational, and governance) was used to provide a predefined coding structure frame. The framework also provided predefined coding categories for critical success factors and value proposition of the implementation (Figure 4 shows this coding structure). Interview transcripts were coded and data attributed to the appropriate category. Challenges, lessons learned, and sustainability were not part of the predefined categories; so new code categories were created. Data were

coded and then checked by multiple researchers for interrater reliability.

All interview data were analyzed to understand the challenges, successes, and lessons learned within each dimension (technical, organizational, and governance) and within each collaborating organization (ConnectVirginia, MEDfx, EHR vendor, and Inova) of on-boarding to an HIE network (in this case ConnectVirginia EXCHANGE). Data were also analyzed to gain insight into issues that would provide meaningful information regarding factors contributing to success, value, and sustainability. Using ATLAS.ti, this was performed by comparing organizations and code families. For example, all stakeholder groups and all codes related to challenges were selected to compare stakeholder positions relative to challenges. Data were then exported to Excel to view frequencies.

Figure 4. Coding structure. CCD=continuity of care document, EHR=electronic health record.



Results

Lessons Learned

The lessons learned, as derived from the challenges and successes, are summarized in Table 3 across technical, organizational, and governance dimensions. Each subsequent subsection (technical, organizational, and governance) serves to unpack those lessons learned in terms of challenges and successes. Collectively, the findings not only provide a retrospective account of Inova's efforts in on-boarding to ConnectVirginia EXCHANGE, but also offer insights into various other stakeholders for future on-boarding efforts.

Table 3. Lessons learned from challenges and successes by dimension (technical, organizational, and governance).

Lessons learned	Challenges	Successes
Technical	Determine the most efficient environment for testing, decoupled from decision processes, actions, and dependencies from other stakeholders.	Willingness to develop workarounds to unexpected software challenges, such as incompatible EHR versions.
	Provide oversight and follow-up to increase technical understanding of appropriate on-boarding guides across all stakeholders.	Gain commitment from implementation site to set high priority on HIE implementation.
	Use an EHR system specific CCD for validation, not a vendor supplied CCD template.	
	Conduct testing and implementation in clearly communicated iterations.	
	Articulate goals and priorities with vendors.	
	Understand roles and required resources in order to minimize time gaps and maximize efficiency.	
Organizational	Account for competing IT priorities across organizations.	Participation of health system leadership.
	Understand, communicate, and appreciate varying stakeholder value proposition/motivations.	Timely and accurate communication, especially by and between the HIE and the health system.
		Allocate appropriate human resources at the outset.
Governance	Ensure governance is in place, including policies, procedures, guidelines, and oversight across all organizations.	Project champion possesses decision-making power, or, as needed delegates appropriate decision making power to others.
	Obtain commitments from governance body early on to facilitate project continuity.	

Technical Dimension

This on-boarding effort between Inova and ConnectVirginia entailed a range of technical activities and coordination to achieve success. There were four major challenges: (1) the testing environment, (2) the on-boarding guides, (3) the CCD, and (4) the vendors, and various successes contributed to the lessons learned.

Challenges

The Testing Environment

Regarding the testing environment, many interviewees felt that significant time early in implementation was spent determining which Inova environment would be used for testing. Inova had two environments capable of producing CCDs: (1) test, and (2) proof of concept (POC); the latter is connected to the EHR vendor's connected network. Inova elected to use the POC environment, which created work inefficiencies across Inova and the EHR vendor. Analysis revealed the challenge, every time Inova wanted to conduct software tests, it required the EHR vendor team to sign into POC, input scenarios, and start the testing. The EHR vendor team considered these manual steps as wasting time, which resulted in testing delays. An interviewee discussed this issue,

> *Had we chosen in the very beginning to use the test, we would not have had any difficulty setting up test patients to test in ConnectVirginia...[New on-boarders] should think long and hard about their testing environment and it should be done early in the process.* [An interviewee]

A review of meeting notes shows that this process took approximately four weeks, whereas participants believed it should have taken less than two weeks. During this time, frustrations from Inova, ConnectVirginia, and MEDfx were observed by one of the researchers, who was consistently on the conference calls. These frustrations were a direct result of Inova's reliance on the vendor related to CCD testing.

The lesson learned from the challenge of choosing a testing environment was that this should be done independent of other stakeholder actions. Determine the most efficient environment for testing, decoupled from decision processes, actions, and other dependencies from other stakeholders.

The On-Boarding Guides

ConnectVirginia provides four guides to assist on-boarders: (1) checklist, (2) implementation, (3) testing, and (4) content, which, unless specified, will be referred to collectively as "on-boarding guides." Not all guides are meant to be read by all parties. For example, the EHR vendor should read content guides, and testing guides should be read by the on-boarding organization. Many of those involved with this implementation had the perspective that the on-boarding guides were not read by the correct parties (or at all by anyone), but that, had they been, certain on-boarding tasks could have gone more smoothly and challenges could have been avoided. For example, the *checklist* is to assist in creating a shared understanding of the data needed, relative to the data available; for example, it requires data on all the laboratory tests that could possibly be run by the on-boarding organization. Several interviewees suggested that this document is long, complex, difficult to read, and

overwhelming to the reader, and therefore does not get completed. However, it is an important element in setting up the CCD for validation. An interviewee suggested that it would be better to go through the checklist section by section to determine whether the data are present or absent.

The *implementation* and *testing* guides are provided to create a shared understanding and level set of expectations about the requirements for implementation and testing. From the meeting notes and interviews, it was clear that these guides were not read to the degree needed for the project. An interviewee explained, "Regardless of how painful [you] think it might be, read the [on-boarding] guides cover to cover." This same interviewee thought the on-boarding guides were very well written and provided plenty of the necessary helpful information. Another interviewee suggested that discussing the on-boarding guides among the implementation group would allow time for conversation and might raise issues not easily discovered by an individual, which are illuminated when discussed in a group. Last, the *content* guide is provided to streamline CCD validation testing. This guide defines what the CCD should have in terms of object identifiers and the corresponding descriptions. Several interviewees felt many misunderstandings could have been avoided had this guide been read and discussed.

The lesson learned was to provide oversight and follow-up to increase technical presentation, reading and understanding of appropriate on-boarding guides across all stakeholders.

Continuity of Care Documents

The challenges with CCDs and HIEs are well documented in the literature [43,45,70]. Initially, the EHR vendor wanted to provide only template CCDs and not CCDs specific to Inova's EHR system, which includes customization. MedVirginia, which conducted the CCD content validation, requested CCDs specific to Inova's system to ensure that a CCD could be sent and received from Inova's customized and nuanced system. Using a standardized vendor template does not account for health system customizations and risks not passing validation in the eHealth Exchange environment. An interviewee noted,

> *A scrubbed [vendor template] CCD will probably not have any issues, but when it comes to testing a CCD from the actual system, there are going to be issues, so you should not expect that just because the sample [vendor template] passed, that the node CCD will pass; it probably won't.* [An interviewee]

Meeting notes and interviews reflect the opinion that this issue required too much time devoted in isolation and should have been handled along with other on-boarding tasks. For example, many participants commented that the technical piece (getting systems to talk) and the content piece (the CCD) should have been conducted in parallel, rather than sequentially. An interviewee thought the CCD issue could have been managed in parallel to solving an issue with handshakes (ie, bidirectional system-to-system acknowledgement),

> *While the CCD is the end point, the handshake had issues. [The Inova] server has to be recognized by MEDfx before any exchange can even happen.* [An interviewee]

Once CCD validation was underway, many felt that the on-boarding process, while excellent and thorough, needed to strike a better balance between the acceptable and the desirable. Several interviewees felt that if all parties had agreed on prioritization of issues (ie, with less attention to details that do not matter), the CCDs could have passed testing much sooner and had an earlier completion date.

The Vendors

For example, the top priority for the EHR vendor was to get the technical pieces to work from their end of the HIE. The vendor's goal was to focus on passing a CCD; not passing a CCD that conforms to the updated eHealth Exchange specifications. Since EHRs undergo a wide variety of customization, this is a critical requirement nuance relative to future participation in eHealth Exchange. While it may have been *acceptable* to use a template CCD, it was *desirable* (because of customizations and future eHealth Exchange participation) to conduct CCD validation with a system CCD. Participants other than the EHR vendor felt the vendor was not focused on developing a technology to support *all* other stakeholder value propositions and motivations.

An interviewee described how differing priorities across stakeholders impacted the HIE outcome,

> *Given the experience with [the EHR vendor], it is important to communicate to clients that this [their decision to limit the CCD] might impact what type of approval is granted at the end of on-boarding, because [Inova] ended up with a conditional approval based on some of the things that we knew [the EHR vendor] wasn't going to budge on.* [An interview]

As noted above, Inova received only conditional approval as an HIE participant/node. The conditional approval was the result of a 90 day medical record date range limitation imposed by the EHR vendor's CCD implementation for this project. The US Social Security Administration requires *more* than three months of medical records in order to conduct disability determinations, but the vendor's implementation allowed only three months' worth of data. In this regard, an interviewee observed,

> *Once we got into the testing process, that unveiled a lot of proprietary issues with [the EHR vendor], even though they say their CCD is compliant. They have certain things that are built into their product, mainly from a competitive stand point. Inova was very reliant on [the EHR vendor], and I think somewhat unaware where those proprietary issues might impact on-boarding to any statewide HIE.* [An interviewee]

Interviewees had comments about both the EHR vendor and MEDfx. Many considered the EHR vendor inflexible and unknowledgeable, especially with regard to the 90 day medical record date range limitation and requirements by the U.S. Social Security Administration, a critical component in the value proposition. Interviewees also felt that MEDfx was developing as the project advanced. According to interviewees, this project represented the first implementation of the latest eHealth Exchange compliant gateway. MEDfx, the gateway provider, had completed development of the gateway in June 2012, but

had not yet completed testing. Thus, Inova became the test bed for the gateway. This process of testing during implementation contributed to the perception that MEDfx was developing as the project was advancing. Both vendors (the EHR vendor and MEDfx) had to fix some bugs that were exposed during testing, and some interviewees thought the fixes should have been completed before implementation or at least done more expeditiously.

The lesson learned was to use an EHR system specific CCD for validation, not a vendor supplied CCD template.

Data gathered from weekly on-boarding meetings, together with document review, reflected the EHR vendor's resistance to meet the needs of a maturing and broader HIE, such as a statewide HIE. An interviewee noted,

> *[The EHR vendor] is a very restricted vendor around allowing third parties to do things like this. We probably lost a month in this go around.* [An Interviewee]

Another interviewee said, "I regret that anyone thought [the EHR vendor] would change their mind."

The lesson learned was that given the complexity of managing expectations across multiple stakeholders, conduct testing and implementation in clearly communicated iterations.

Several other challenges surfaced regarding software versions that were incompatible with the latest data exchange standards. The EHR vendor software version that Inova used for this implementation (released in 2010) was not compliant with the upcoming eHealth Exchange specifications. The vendor will not release a version compliant with the new specifications until the 2012 version. Discussions with the EHR vendor suggested that, due to the relative newness of the current 2010 version, health systems will likely not deploy the 2012 version for some time. This discrepancy in EHR specifications created significant challenges, in terms of the CCD content, to completing on-boarding. However, the EHR vendor's perspective differed from that of ConnectVirginia regarding the ability to on-board, saying, "We have many other clients that have on-boarded to eHealth Exchange, and none of them have these [CCD content] issues that ConnectVirginia is citing." In response, ConnectVirginia participants explained that the clients to which the EHR vendor refers had been on-boarded under the old Nationwide Health Information Network specifications established by ONC, rather than the new and required eHealth Exchange standards established in September 2012. ConnectVirginia must on-board to eHealth Exchange under the updated eHealth Exchange specifications. Thus, misunderstandings about the required specifications caused significant implementation delays. Because this was the first on-boarding with the EHR vendor, there were many unknowns. It became apparent that an early meeting with the EHR vendor was critical. Many felt this would have created a shared understanding of some of the nuances of each other's systems and of stakeholders' motivations, while also fostering conversations about how each system adheres to the implementation specifications.

The lesson learned was to establish early meetings with vendors to articulate goals and priorities.

Interviews revealed challenges with understanding each stakeholder's roles. Several interviewees felt that time was lost determining who was responsible for certain things, and tasks were not done because one person thought another was responsible. Likewise, better understanding was needed of the technical resources available: (1) Are the right people working in the right place?; and (2) Is the testing environment one that will facilitate on-demand testing?. An interviewee felt that two MEDfx people had the knowledge collectively, but lacked depth individually. This type of situation led to delays or multiple attempts to get questions answered. Another interviewee felt it was critical to have representation across integrated delivery teams, primarily because policy issues needed to be addressed saying, "An interface group will build a pipe for your data to pass, but there are lots of rules regarding audit streams." A majority of interviewees thought many of these late questions or realizations could have been avoided by earlier and better understanding of the on-boarding guides provided to Inova and the EHR vendor. Much that was done was conducted sequentially; many interviewees thought the technical piece (getting the systems to talk) and the content piece (the CCD) should have been done in parallel and speculated that doing so would have saved a lot of time.

The lesson learned was to conduct clear communication early on to discuss and understand roles and required resources in order to minimize time gaps and maximize efficiency.

Successful Software Redevelopment

Regarding the incompatible software versions described above with the 2010 EHR version, almost all the interviewees with knowledge of this issue commented on MEDfx's willingness to develop new code to address this challenge. While software redevelopment took time, causing unanticipated delays, everyone saw this as a significant success. A participant summarized the thoughts of many,

> *We put a lot of responsibility on MEDfx to make adjustments on their side to accommodate the fact that [Inova] was running a version that only supported 2010. Thankfully, they were flexible enough to accommodate that.* [A participant]

While Inova could have on-boarded to ConnectVirginia without this modification, ConnectVirginia would not have been able to on-board to eHealth Exchange, thus minimizing the value for Inova and any organization on-boarding after Inova. All interviewees felt that getting Inova on-boarded was a great success in and of itself.

All stakeholders had competing priorities, but participants noted that Inova's may been the most significant. Although they were in the middle of an EHR implementation, they chose to pioneer on-boarding to the ConnectVirginia EXCHANGE.

The lesson learned was to gain commitment from technology stakeholders to be willing to develop workarounds to unexpected software challenges, such as incompatible EHR versions.

Another lesson learned was to determine conflicting priorities across stakeholders at the outset. Gain commitment from implementation site to set high priority on HIE implementation.

Organizational Dimension

Organizational factors were also instrumental to the success of the Inova on-boarding experience. Concurrent EHR implementation and strong leadership contributed to the organizational challenges, successes, and lessons learned.

Challenging Competing Priorities

The major challenge involved a concurrent EHR implementation at Inova and understanding each stakeholder's value proposition and motivation. Early on, the concurrent EHR implementation created a situation of competing priorities. However, once roles were more clearly defined regarding the ConnectVirginia implementation, it was felt that resources were available and engaged. As one interviewee observed, "[The Inova internal champion] kept us [Inova team] moving because we were very busy with a lot of other stuff including [EHR] implementation."

The lesson learned was to account for major competing IT priorities at each participating organization. A concurrent EHR implementation will likely compete directly with the HIE implementation.

The value proposition and corresponding motivation for on-boarding to ConnectVirginia EXCHANGE varied with the stakeholder. Several Inova participants commented that, although the initial economic value proposition to Inova was nonexistent, the motivation to move forward was very well aligned with their vision to "reinvent hospital-based care to increase value for our patients" and to "look outside our hospitals to build an integrated network of providers and programs to support our community." The culture of this vision was embedded in Inova employee beliefs. The words of several were summed up by one Inova interviewee, "On-boarding to ConnectVirginia [exchange] aligns with the Inova vision, fulfills our desire to be part of transforming the Commonwealth [of Virginia] into a great place to be a patient, and success for Inova means great benefit for the community."

Additional motivations for Inova involved the desire to be leaders in the HIE trend. Some interviewees questioned whether or not the EHR vendor understood why this was so important to Inova and ConnectVirginia independently and collectively, and interviewees suggested the EHR vendor was sometimes argumentative with requests from the on-boarding team, "Their [the EHR vendor's] resistance to ConnectVirginia's success is what concerns me." To the other stakeholders, it seemed that the EHR vendor did not have a clear motivation and was simply responding reluctantly to client requests. These differing views on value created communication breakdowns, frustrations, and inefficiencies in the on-boarding process. These breakdowns and frustrations were observed numerous times on various conference calls with the implementation team. There were times when it would take three or four calls to resolve one issue. Such events led to inefficiencies in the on-boarding process.

The lesson learned was to understand, communicate, and appreciate varying stakeholder value proposition/motivations.

Successful Leadership

Leadership was an important factor in this on-boarding process. Almost all interviewees commented that Inova's internal champion, the Executive Vice President and Chief Technology Officer of Inova, was a critical component in the success of the project. Many suggested that his role on the ConnectVirginia Governing Body put him in a position to be an internal champion not only for Inova, but for ConnectVirginia as well, with his solid understanding of what ConnectVirginia was trying to accomplish and why it was important. During an interview, he stated,

> *I thought Inova needed to learn what HIE is and needed to get its feet wet with the on-boarding so it could be connected. I believe in HIE.* [Executive Vice President and Chief Technology Officer of Inova]

Scheduling sensitivities around Inova's EHR implementation created some time periods when key people were unavailable. When this resulted in a lack of progress between meetings, the Executive Vice President and Chief Technology Officer of Inova could help guide the Inova team with managing those competing priorities. His unique combination of being an internal champion and a decision-maker greatly enhanced the success of this project. An interviewee further qualified the role of an internal champion, "This cannot be a technical champion, but a true champion...a true leader." Leadership in terms of project management was considered solid. Several interviewees commented on Inova's PM, and one summarized the words of many,

> *She [the PM] was prepared, answered emails promptly and completely, and executed well. It was extremely helpful to have her.* [Interviewee]

The lesson learned was that the participation of health system leadership is critical to success.

Communication was another area of success, and many felt that PMs from both Inova and ConnectVirginia greatly contributed to that success. As mentioned above, the PM from Inova was always prepared and answered emails promptly, and many others commented on the PM from ConnectVirginia. An interviewee capsulized ConnectVirginia's communication efforts,

> *I appreciate the fact that they controlled a lot of the documentation. They scheduled the weekly calls, set up the agendas, sent out the minutes, and managed any outstanding items across everyone.* [Interviewee]

Several interviewees mentioned that the meeting minutes were very thorough.

The lesson learned was that timely and accurate communication, especially by and between the HIE and health system, is essential.

Despite the fact that this project competed with resources for the Inova EHR implementation, many felt there were appropriate resources. However, most interviewees observed that initially the correct resources were not allocated. Since this was the first on-boarding, there were vague expectations about the level of and appropriateness of resources. Teams required skills sets,

knowledge, and experience that were not available at the outset. An interviewee noted,

> *Too many assumptions were made. We need to have a better kick-off to level set expectations and roles.* [An Interviewee]

Most interviewees felt that, by the second month, the teams were appropriately resourced, and what was originally a challenge became a success. Several commented that there was not a lot of movement regarding the resources, which added to the teams' strength individually and collectively. Another interviewee summed up the lesson learned,

> *Put your best resources around standing this up, because it requires you to pay attention to detail. This is more than an IT project; this is not a simple interface project.* [An Interviewee]

The lesson learned was to allocate appropriate human resources at the outset.

Governance Dimension

Intra and interorganizational decision-making power and clear role definition have been shown to decrease intra and interorganizational issues [71]. Furthermore, the nature of the relationships between decision makers is important in navigating variable governance processes and structures and in sharing decisions. Governance is the establishment of oversight, standardized policies and procedures, and mechanisms to ensure operation of an organization [72]. Thus, governance factors such as the on-boarding guides, project resources, and a project champion contributed to the lessons learned and the critical success factors of the Inova on-boarding project.

Challenging Identification of Appropriate Policies and Guidelines

Identifying the appropriate policies and guidelines across stakeholders was challenging, as was selecting the best people to provide oversight. Unfortunately, a structured governance process did not predate this project, this was the first time this particular group of organizations had worked together, and the first time on-boarding to ConnectVirginia had taken place. As noted previously, providing oversight and policy enforcement for people to read the on-boarding guides proved challenging. If on-boarding guides had been read thoroughly, it may have been easier to identify the correct governance resources or, at least, ask questions regarding resource selection. Regarding the governance structure for the project, one interviewee commented, "The technology supports the business, but the business does not go anywhere without the right folks." Another interviewee said that she was, "...challenged to put together a governance group that could attend the weekly on-boarding meetings, as those were a great way of getting decisions made."

These deficits resulted in the organizational and technical challenges described earlier, including missing nuances of the on-boarding process, difficulties selecting key project resources at the outset of the project, and not providing a structure whereby team members could request guidance in resource selection. Fortunately, these issues were identified and quickly rectified within the first two months of the project.

The lesson learned was to ensure governance is in place, including policies, guidelines, and oversight across all organizations.

Most interviewees agreed that the on-boarding guides provided by ConnectVirginia were useful in explaining appropriate governance such as policies, procedures, etc, once they were read. The challenge was getting people to read them. Time was short, priorities competitive, and resources thin. However, several interviewees agreed that thorough reading of the on-boarding guides just before the kick-off meeting would have helped ensure that proper governance decisions were made, especially in regards to policy decisions. It was also thought that a thorough reading would have mitigated some downstream misunderstandings and poor understanding of the system requirements. Other than that, some interviewees thought the ConnectVirginia PM could have done more to ensure that those responsible for governance understood the on-boarding guides pertaining to their part of the project. For example, at one on-boarding meeting, one individual from Inova asked to go through one of the on-boarding guides. An interviewee commented, "Sometimes we all need to have our hand held, and if that is what it would have taken to make sure everyone went through the on-boarding guides, then so be it."

The lesson learned was to obtain commitments from the governance body early on to facilitate project continuity.

Successful Project Champion

Most interviewees agreed that the real success in this project, from a governance perspective, was having a project champion with decision-making power. Several times policy or procedure decisions needed to be made; and because the Executive Vice President and Chief Technology Officer of Inova was involved, the appropriate questions were asked and decisions made. An interviewee gave a good example of how the project champion provided appropriate decision-making authority,

> *When something comes up as an issue, helping to figure out if it is a technical issue or a policy issue. Then figuring out whose issue it is; is it a node [hospital system] issue, is it the vendor, or is it ConnectVirginia?* [An Interviewee]

In these situations, the Executive Vice President and Chief Technology Officer of Inova was able to provide guidance on issues involving the node or Inova's EHR vendor. Regarding MEDfx and issues attributed to them, the PM had authority to provide the guidance needed to move forward. Regarding CCD content validation, which was conducted by MedVirginia, it was felt that the person on the on-boarding calls did not have decision-making authority, and thus needed to seek guidance after the call. Yet, interviewees felt her follow-up communications were timely and comprehensive.

The lesson learned was that a project champion is essential who possesses decision-making power, or, as needed, delegates appropriate decision-making power to others.

It was clear to participants how critical it was to ensure from the beginning that a proper oversight structure is in place, including the people involved in the project. As mentioned in

the challenges section, the on-boarding guides provided by ConnectVirginia helped to provide guidance in this regard and to identify a governance structure. In addition to identifying project resources, early identification of an internal champion is essential for project success. But, as one interviewee mentioned, it is sometimes difficult to have the internal champion with decision-making authority at the weekly meetings.

Stakeholder Perceived Value

Analysis of interviews, observations, and project documents, taken together, also revealed a crosscutting theme in terms of the goals, priorities, motivations, and perceived value of engaging in HIE. Earlier on in the implementation, an important success factor would have been to have a better understanding of how well the organizational goals of each participating organization aligned with one another; the implementation priorities and motivations (social and economic) for each organization to participate; and the perceived value that each organization expected to gain as a result of participating. These are illustrated in Table 4. Providing this information may have avoided some of the challenges, constraints, and tensions experienced, especially with the EHR vendor.

It was felt by many interviewees that had something like the below matrix existed, that clarity and insight would have been gained early on.

Table 4. Matrix of goals, priorities, motivations, and perceived value propositions across implementation stakeholders.

	ConnectVirginia	Inova	MEDfx	EHR vendor
Implementation goal alignment	Aligned	Aligned	Aligned	Not aligned
Implementation priority	High	High	High	Low
Motivation	Social high	Social high	Social low	Social low
	Economic moderate	Economic low	Economic high	Economic low
Perceived value	Provides exchange of medical information	HIE leader in the state medical information at the point of care	Fulfills the contract terms	None apparent
	Fulfills the contract terms	Social Security Administration disability determination		

Discussion

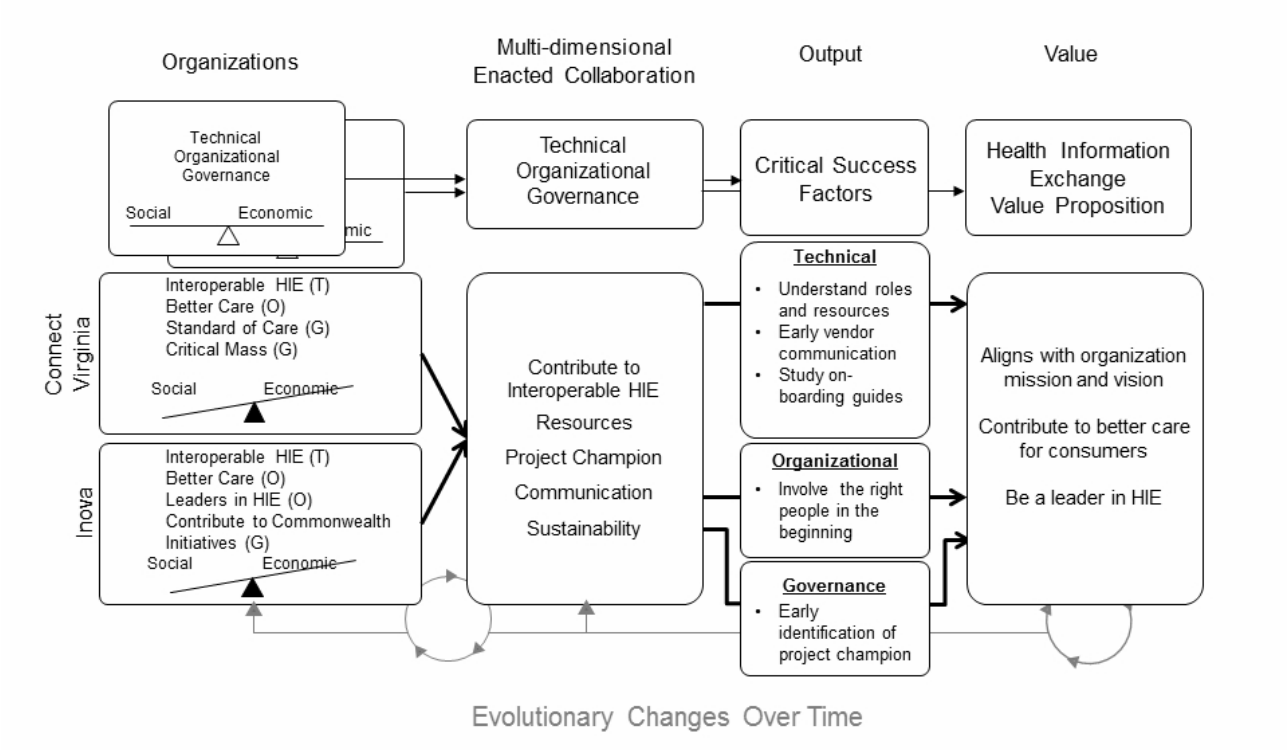
Principal Findings

The main findings of this case study included several challenges, successes, and lessons learned during the implementation of a health system on-boarding to a statewide HIE. Figure 5 shows a summary of the Blended Value Collaboration Enactment Framework as applied to these findings and illustrates the technical, organizational, and governance collaboration that took place between ConnectVirginia and Inova, along with critical success factors and associated value proposition details.

This case study demonstrates that *interorganizational governance* of HIE implementation is replete with interrelated and overlapping technical, organizational, and governance issues. The complexities of collaboration appear to assist as well as detract from realizing a set of common goals. For the expanded view of HIE (ie, across states and the nation), the broader significance of this case study is the proposition that successful implementation of a large-scale emergent HIE system should consider the expected and realized blended value across all participants. Consistent with the literature, while economic value is an important goal, the organizations presented here have regarded the value proposition primarily to include social value, believing that economic value will follow at some point in time.

As mentioned earlier, the sociotechnical approach allows for understanding independent linkages between complex social and technological components. Much of what was learned from this first on-boarding effort is related to accomplishing tasks earlier in the process, rather than allowing them to be discovered in course. While this may be common knowledge in more established IT implementations, the field of statewide HIE on-boarding implementations is undeveloped in this area. The interviews illuminated some common issues such as interoperable HIE and better care. However, it was notable that both organizations, when asked about their motivation to collaborate, cited the social good that would result from creating a critical mass and contributing to the Commonwealth's HIE initiative. Since Inova has no other organizations with which to exchange, and ConnectVirginia does not yet receive revenue from on-boarding organizations, both organizations were motivated primarily for the social good.

Figure 5. Blended Value Collaboration Enactment Framework as applied to findings. T=technical, O=organizational, G=governance, and HIE=health information exchange.



Application of the Framework

Turning to the Blended Value Collaboration Framework as applied to the findings, Figure 5 illustrates the interdependent linkages between the technical, organizational, and governance dimensions for each organization. When these individual dimensions come together in the collaboration there are contributing factors that facilitate enactment of the collaboration. At this early point in the project, both organizations were heavily weighted toward social motivations, recognizing that as ConnectVirginia matures the economic motivation will grow more prominent.

Frequently in public-private collaborations, the value propositions of collaborating organizations are not aligned. However, in this case, we found the value propositions between ConnectVirginia and Inova to be very well aligned and centered on organizational missions and goals, better care for consumers, and being leaders in HIE. This is very consistent with the IOS literature [41,42]. As mentioned earlier, this study focused on the implementation, so there was no usage, and only one period of time was studied. Thus, the evolutionary changes over time portion of the framework were not addressed in these findings (gray portion in Figure 5). This sociotechnical approach facilitates the consideration of the social and the technical perspectives, and their contribution to the overall value proposition.

Supporting Mody's [32] assertion, describer earlier in this paper, that social considerations could exert more pressure than economic considerations, this framework highlights social reasons as very significant motivators for early adopters/implementers, frequently outweighing economic motivations. Lacking a framework that considers social motivations, the natural tendency in IT projects might be to analyze, or only consider, economic motivators, and then judge the value proposition on that basis. The Blended Value Collaboration Enactment Framework fills a gap in, and makes a contribution to, the STS framework literature. Its application promotes the assessment of a wide range of observed issues (across technical, organizational, and governance dimensions) in relation to an interorganizational health IT implementation, comparing them with IOS goals, motivations, and intraorganizational priorities, and then determining the success factors and value propositions from the results. Used as a heuristic, the framework may provide for a broader and more inclusive evaluation of an IOS health IT implementation. Furthermore, the current framework enables examination of changing motivations and value propositions over time.

Future studies should revisit the findings reported here to analyze such changes. As time progresses and ConnectVirginia matures, an increased economic motivation is expected from both organizations. In the future, organizations are also expected to on-board primarily for perceived potential economic factors, although these are yet to be realized. Future research should assess a wide range of economic and clinical factors associated with HIE value; while continuing to define, include, and broaden social factors and public value. Mixed-method case studies can be used in this regard to more fully understand the breadth and depth of mediating, moderating, and control variables to assess in future quantitative studies. Studying HIE implementations broadly across the United States through survey research is desperately needed as most studies, including those referenced in this manuscript, consist of small sample sizes. In terms of content, the value proposition in HIE is a moving target, as both the act of HIE and entity called HIE continues to evolve and

change. New government requirements and incentives, new business models to facilitate HIE, and increased societal demand for better and less expensive health care are expected to continue to shape the HIE landscape. Knowing this evolution will occur should not deter near-term research. These studies are needed in order to make the ongoing practical impact discussed at the outset of this manuscript, to address the triple aim of health care broadly across regions.

Limitations

This case study examines the implementation of one health system on-boarding to a statewide HIE. As such, generalizability may be limited. Another factor holding potential to contribute to this limitation includes that the Chief Medical Information Office of the health system was the statewide HIE Governing Body Vice-Chair, which may have served to introduce motivations and bias that a different implementation would not have experienced. However, it may also be commonplace for existing health care leaders in regions and states to take a strong role in HIE governance. Further research is needed to apply the principles from this study to other implementations, so as to gain generalizability of the findings.

Conclusions

This study focuses on the evaluation of HIE implementation in Virginia. From a practical perspective, the study provides a set of lessons learned for others who are implementing systems across a statewide HIE. This study also includes considerations for eHealth Exchange implementation. As mentioned earlier and substantiated in the literature, on-boarding to eHealth Exchange is part of the economic value proposition equation. On-boarding to ConnectVirginia with a CCD that will not pass testing when ConnectVirginia on-boards to eHealth Exchange eliminates a critical value proposition component. From a methodological perspective, it provides an example of how such an HIE implementation can be studied, and from a theoretical perspective, this study builds on the literature on IOS for health care, addressing the core questions: (1) What value propositions motivate an organization to participate in HIE implementation?; and (2) What success factors should be targeted in HIE implementation evaluation?.

Acknowledgments

The authors acknowledge and thank the interviewees who participated in this study. Their time was very much appreciated. Their various perspectives provided valuable input to this case study and useful information for other organizations in the on-boarding process. The authors also thank the reviewers whose suggestions made this manuscript more valuable.

Conflicts of Interest

While this case study was being conducted, the corresponding author was contracted as the external evaluator, a federal cooperative agreement requirement, to CHA, the subcontractor responsible for building ConnectVirginia. There are no other real or perceived conflicts of interest.

References

1. State health information exchange. State health information exchange cooperative agreement program URL: http://www.healthit.gov/policy-researchers-implementers/state-health-information-exchange [accessed 2014-04-03] [WebCite Cache ID 6OYsOqD8R]
2. The IHI triple aim. URL: http://www.ihi.org/engage/initiatives/TripleAim/Pages/default.aspx [accessed 2014-04-03] [WebCite Cache ID 6OYsd691v]
3. Furukawa MF, Patel V, Charles D, Swain M, Mostashari F. Hospital electronic health information exchange grew substantially in 2008-2012. Health Aff (Millwood) 2013 Aug;32(8):1346-1354. [doi: 10.1377/hlthaff.2013.0010] [Medline: 23918477]
4. Vest JR, Gamm LD. Health information exchange: Persistent challenges and new strategies. J Am Med Inform Assoc 2010;17(3):288-294 [FREE Full text] [doi: 10.1136/jamia.2010.003673] [Medline: 20442146]
5. Unertl KM, Johnson KB, Gadd CS, Lorenzi NM. Bridging organizational divides in health care: An ecological view of health information exchange. JMIR Med Inform 2013 Oct 29;1(2):e3. [doi: 10.2196/medinform.2510]
6. Shade SB, Chakravarty D, Koester KA, Steward WT, Myers JJ. Health information exchange interventions can enhance quality and continuity of HIV care. Int J Med Inform 2012 Oct;81(10):e1-e9. [doi: 10.1016/j.ijmedinf.2012.07.003] [Medline: 22854158]
7. Vest JR, Miller TR. The association between health information exchange and measures of patient satisfaction. Appl Clin Inform 2011;2(4):447-459 [FREE Full text] [doi: 10.4338/ACI-2011-06-RA-0040] [Medline: 23616887]
8. McGee MK. InformationWeek. 2010. Health information enhances decision making URL: http://www.informationweek.com/healthcare/clinical-information-systems/health-information-exchange-enhances-decision-making/d/d-id/1090004?page_number=1 [accessed 2014-04-03] [WebCite Cache ID 6OYsvJgCx]
9. Dimitropoulos L, Patel V, Scheffler SA, Posnack S. Public attitudes toward health information exchange: Perceived benefits and concerns. Am J Manag Care 2011 Dec;17(12 Spec No):SP111-SP116 [FREE Full text] [Medline: 22216769]
10. Hincapie A, Warholak T. The impact of health information exchange on health outcomes. Appl Clin Inform 2011;2(4):499-507 [FREE Full text] [doi: 10.4338/ACI-2011-05-R-0027] [Medline: 23616891]

11. Frisse ME, Johnson KB, Nian H, Davison CL, Gadd CS, Unertl KM, et al. The financial impact of health information exchange on emergency department care. J Am Med Inform Assoc 2012;19(3):328-333 [FREE Full text] [doi: 10.1136/amiajnl-2011-000394] [Medline: 22058169]
12. Kaelber DC, Bates DW. Health information exchange and patient safety. J Biomed Inform 2007 Dec;40(6 Suppl):S40-S45 [FREE Full text] [doi: 10.1016/j.jbi.2007.08.011] [Medline: 17950041]
13. Hessler BJ, Soper P, Bondy J, Hanes P, Davidson A. Assessing the relationship between health information exchanges and public health agencies. J Public Health Manag Pract 2009;15(5):416-424. [doi: 10.1097/01.PHH.0000359636.63529.74] [Medline: 19704310]
14. Dobbs D, Trebatoski M, Revere D. The northwest public health information exchange's accomplishments in connecting a health information exchange with public health. Online J Public Health Inform 2010;2(2) [FREE Full text] [doi: 10.5210/ojphi.v2i2.3210] [Medline: 23569585]
15. Shapiro JS, Mostashari F, Hripcsak G, Soulakis N, Kuperman G. Using health information exchange to improve public health. Am J Public Health 2011 Apr;101(4):616-623. [doi: 10.2105/AJPH.2008.158980] [Medline: 21330598]
16. Committee on Quality of Health Care in America, Institute of Medicine. Crossing the quality chasm: A new health system for the 21st century. Washington, D.C: National Academy Press; 2001.
17. Fontaine P, Ross SE, Zink T, Schilling LM. Systematic review of health information exchange in primary care practices. J Am Board Fam Med 2010;23(5):655-670 [FREE Full text] [doi: 10.3122/jabfm.2010.05.090192] [Medline: 20823361]
18. Bailey JE, Wan JY, Mabry LM, Landy SH, Pope RA, Waters TM, et al. Does health information exchange reduce unnecessary neuroimaging and improve quality of headache care in the emergency department? J Gen Intern Med 2013 Feb;28(2):176-183 [FREE Full text] [doi: 10.1007/s11606-012-2092-7] [Medline: 22648609]
19. Brailer DJ. Interoperability: The key to the future health care system. Health Aff (Millwood) 2005;Suppl Web Exclusives:W5-19 [FREE Full text] [doi: 10.1377/hlthaff.w5.19] [Medline: 15659454]
20. Kuperman GJ. Health-information exchange: Why are we doing it, and what are we doing? J Am Med Inform Assoc 2011;18(5):678-682 [FREE Full text] [doi: 10.1136/amiajnl-2010-000021] [Medline: 21676940]
21. Unertl KM, Johnson KB, Lorenzi NM. Health information exchange technology on the front lines of healthcare: Workflow factors and patterns of use. J Am Med Inform Assoc 2012;19(3):392-400 [FREE Full text] [doi: 10.1136/amiajnl-2011-000432] [Medline: 22003156]
22. Miller RH, Miller BS. The Santa Barbara County care data exchange: What happened? Health Aff (Millwood) 2007;26(5):w568-w580 [FREE Full text] [doi: 10.1377/hlthaff.26.5.w568] [Medline: 17670775]
23. California Healthcare Foundation. Achieving the right balance: Privacy and security policies to support electronic health information exchange. 2012 URL: http://www.chcf.org/~/media/MEDIA%20LIBRARY%20Files/PDF/A/PDF%20AcheivingBalancePrivacySecurityHIE.pdf [accessed 2014-04-03] [WebCite Cache ID 6OYtRTlwS]
24. Adler-Milstein J, McAfee AP, Bates DW, Jha AK. The state of regional health information organizations: Current activities and financing. Health Aff (Millwood) 2008;27(1):w60-w69 [FREE Full text] [doi: 10.1377/hlthaff.27.1.w60] [Medline: 18073225]
25. McDonald CJ, Overhage JM, Barnes M, Schadow G, Blevins L, Dexter PR, INPC Management Committee. The Indiana network for patient care: A working local health information infrastructure. An example of a working infrastructure collaboration that links data from five health systems and hundreds of millions of entries. Health Aff (Millwood) 2005;24(5):1214-1220 [FREE Full text] [doi: 10.1377/hlthaff.24.5.1214] [Medline: 16162565]
26. Stead WW, Kelly BJ, Kolodner RM. Achievable steps toward building a national health information infrastructure in the United States. J Am Med Inform Assoc 2005;12(2):113-120 [FREE Full text] [doi: 10.1197/jamia.M1685] [Medline: 15561783]
27. Walker J, Pan E, Johnston D, Adler-Milstein J, Bates DW, Middleton B. The value of health care information exchange and interoperability. Health Aff (Millwood) 2005;Suppl Web Exclusives:W5-10 [FREE Full text] [doi: 10.1377/hlthaff.w5.10] [Medline: 15659453]
28. Whitworth B, Ahmad A, Soegaard M, Dam RF. The encyclopedia of human-computer interaction. 2nd ed. Aarhus, Denmark: The Interaction Design Foundation; 2013. Socio-technical system design URL: http://www.interaction-design.org/encyclopedia/socio-technical_system_design.html [accessed 2014-04-03] [WebCite Cache ID 6OYtl5VWu]
29. Porter M, Teisberg E. Creating value-based competition on results. In: Redefining health care. USA: Harvard Business School Press; 2006.
30. Bowen PL, Cheung MD, Rohde FH. Enhancing IT governance practices: A model and case study of an organization's efforts. International Journal of Accounting Information Systems 2007 Sep;8(3):191-221. [doi: 10.1016/j.accinf.2007.07.002]
31. Reich BH, Benbasat I. An empirical investigation of factors influencing the success of customer-oriented strategic systems. Information Systems Research 1990 Sep;1(3):325-347. [doi: 10.1287/isre.1.3.325]
32. Mody A. Infrastructure delivery: Private initiative and the public good. Washington, D.C: World Bank; 1996.
33. Emerson J. California Management Review. 2003. The blended value proposition: Integrating social and financial returns URL: http://www.blendedvalue.org/the-blended-value-proposition-integrating-social-and-financial-returns/ [accessed 2014-06-08] [WebCite Cache ID 6QBGzxC7N]

34. Whitworth B. The social requirements of technical systems. In: Moor AD, editor. Handbook of research on socio-technical design and social networking systems (2-volumes). USA: Information Science Reference; 2009.
35. Cash JI, Kosynski BR. Harv Bus Rev. 1985 Apr. IS redraws competitive boundaries URL: http://hbr.org/1985/03/is-redraws-competitive-boundaries/ar/1 [accessed 2014-07-28] [WebCite Cache ID 6RPGlR1oq]
36. Williams CB, Federowicz J. A framework for analyzing cross-boundary e-government projects: The CapWin example. : Proceedings of the 2005 National Conference on Digital Government Research, DG. O; 2005 Presented at: USA; May 2005; Atlanta, Georgia, USA p. 15-18 URL: http://dl.acm.org/citation.cfm?id=1065337&dl=ACM&coll=DL&CFID=516585855&CFTOKEN=36368055
37. Premkumar G, Ramamurthy K. The role of interorganizational and organizational factors on the decision mode for adoption of interorganizational systems. Decision Sciences 1995 May;26(3):303-336. [doi: 10.1111/j.1540-5915.1995.tb01431.x]
38. Melville N, Kraemer K, Gurbaxani V. MIS Quarterly. 2004. Review: Information technology and organizational performance: An integrative model of IT business value URL: http://misq.org/review-information-technology-and-organizational-performance-an-integrative-model-of-it-business-value.html [accessed 2014-07-28] [WebCite Cache ID 6RPHHC9DK]
39. Barua A, Konana P, Whinston AB, Yin F. MIS Quarterly. 2004. An empirical investigation of net-enabled business value URL: http://misq.org/an-empirical-investigation-of-net-enabled-business-value.html [accessed 2014-07-28] [WebCite Cache ID 6RPHR7DeR]
40. Delone WH, McLean ER. Journal of Management Information Systems. 2003. The DeLone and McLean model of information systems success: A ten-year update URL: http://www.mesharpe.com/MISVirtual/07Delone.pdf [accessed 2014-07-28] [WebCite Cache ID 6RPHZlKCg]
41. Bardach E. Getting agencies to work together: The practice and theory of managerial craftmanship. Washington, D.C: Brookings Institution Press; 1998.
42. Yusof MM, Papazafeiropoulou A, Paul RJ, Stergioulas LK. Investigating evaluation frameworks for health information systems. Int J Med Inform 2008 Jun;77(6):377-385. [doi: 10.1016/j.ijmedinf.2007.08.004] [Medline: 17904898]
43. Feldman SS, Horan TA. Collaboration in electronic medical evidence development: A case study of the Social Security Administration's MEGAHIT System. Int J Med Inform 2011 Aug;80(8):e127-e140. [doi: 10.1016/j.ijmedinf.2011.01.012] [Medline: 21333588]
44. Saef S, Bourne C, Bush J, Scott L, Gaafary H, Keenan K, et al. The impact of a health information exchange on resource use and medicare-allowable charges at eleven emergency departments operated by four major hospital systems in a midsized Southeastern city: An observational study using clinician estimates. Annals of Emergency Medicine 2013 Oct;62(4):S97. [doi: 10.1016/j.annemergmed.2013.07.090]
45. Feldman SS. Public-private interorganizational sharing of health data for disability determination. Ann Arbor, MI: ProQuest, UMI; 2011.
46. Inova overview. URL: http://www.inova.org/about-inova/index.jsp [accessed 2014-04-03] [WebCite Cache ID 6OYuGVJGe]
47. Ash JS, Guappone KP. Qualitative evaluation of health information exchange efforts. J Biomed Inform 2007 Dec;40(6 Suppl):S33-S39 [FREE Full text] [doi: 10.1016/j.jbi.2007.08.001] [Medline: 17904914]
48. Johnson KB, Gadd C. Playing smallball: Approaches to evaluating pilot health information exchange systems. J Biomed Inform 2007 Dec;40(6 Suppl):S21-S26 [FREE Full text] [doi: 10.1016/j.jbi.2007.08.006] [Medline: 17931981]
49. Cooper RB, Zmud RW. Information technology implementation research: A technological diffusion approach. Management Science 1990 Feb;36(2):123-139. [doi: 10.1287/mnsc.36.2.123]
50. Hripcsak G, Kaushal R, Johnson KB, Ash JS, Bates DW, Block R, et al. The United Hospital Fund meeting on evaluating health information exchange. J Biomed Inform 2007 Dec;40(6 Suppl):S3-10 [FREE Full text] [doi: 10.1016/j.jbi.2007.08.002] [Medline: 17919986]
51. Ammenwerth E, Gräber S, Herrmann G, Bürkle T, König J. Evaluation of health information systems-problems and challenges. Int J Med Inform 2003 Sep;71(2-3):125-135. [Medline: 14519405]
52. Marchibroda JM. Health information exchange policy and evaluation. J Biomed Inform 2007 Dec;40(6 Suppl):S11-S16 [FREE Full text] [doi: 10.1016/j.jbi.2007.08.008] [Medline: 17981099]
53. Kern LM, Kaushal R. Health information technology and health information exchange in New York State: New initiatives in implementation and evaluation. J Biomed Inform 2007 Dec;40(6 Suppl):S17-S20 [FREE Full text] [doi: 10.1016/j.jbi.2007.08.010] [Medline: 17945542]
54. Labkoff SE, Yasnoff WA. A framework for systematic evaluation of health information infrastructure progress in communities. J Biomed Inform 2007 Apr;40(2):100-105 [FREE Full text] [doi: 10.1016/j.jbi.2006.01.002] [Medline: 16530489]
55. Vest JR, Jasperson '. How are health professionals using health information exchange systems? Measuring usage for evaluation and system improvement. J Med Syst 2012 Oct;36(5):3195-3204 [FREE Full text] [doi: 10.1007/s10916-011-9810-2] [Medline: 22127521]
56. Schooley BL, Horan TA. Towards end-to-end government performance management: Case study of interorganizational information integration in emergency medical services (EMS). Government Information Quarterly 2007 Oct;24(4):755-784. [doi: 10.1016/j.giq.2007.04.001]

57. Glasgow RE. eHealth evaluation and dissemination research. Am J Prev Med 2007 May;32(5 Suppl):S119-S126. [doi: 10.1016/j.amepre.2007.01.023] [Medline: 17466816]
58. McGowan JJ, Cusack CM, Poon EG. Formative evaluation: A critical component in EHR implementation. J Am Med Inform Assoc 2008;15(3):297-301 [FREE Full text] [doi: 10.1197/jamia.M2584] [Medline: 18308984]
59. Weick KE. Small wins: Redefining the scale of social problems. American Psychologist 1984;39(1):40-49. [doi: 10.1037/0003-066X.39.1.40]
60. Fountain JE. Building the virtual state: Information technology and institutional change. Washington, D.C: Brookings Institution Press; 2001.
61. Geels FW. From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from sociology and institutional theory. Research Policy 2004 Sep;33(6-7):897-920. [doi: 10.1016/j.respol.2004.01.015]
62. Kern LM, Barron Y, Abramson EL, Patel V, Kaushal R. HEAL NY: Promoting interoperable health information technology in New York State. Health Aff (Millwood) 2009;28(2):493-504 [FREE Full text] [doi: 10.1377/hlthaff.28.2.493] [Medline: 19276009]
63. Markus ML. Power, politics, and MIS implementation. Commun. ACM 1983;26(6):430-444. [doi: 10.1145/358141.358148]
64. Ash JS, Anderson NR, Tarczy-Hornoch P. People and organizational issues in research systems implementation. J Am Med Inform Assoc 2008;15(3):283-289 [FREE Full text] [doi: 10.1197/jamia.M2582] [Medline: 18308986]
65. Kuperman GJ. Doing interdisciplinarity: Motivation and collaboration in research for sustainable agriculture in the UK. J Am Med Inform Assoc 2011;18(5):678-682 [FREE Full text] [doi: 10.1111/j.1475-4762.2008.00859.x]
66. Ramírez R. Value co-production: Intellectual origins and implications for practice and research. Strat. Mgmt. J 1999 Jan;20(1):49-65 [FREE Full text] [doi: 10.1002/(SICI)1097-0266(199901)20:1<49::AID-SMJ20>3.0.CO;2-2]
67. Bititci US, Martinez V, Albores P, Parung J. Creating and managing value in collaborative networks. International Journal of Physical Distribution & Logistics Management 2004;34(3):251-268. [doi: 10.1108/09600030410533574]
68. Friesen ME, Johnson JA. The success paradigm: Creating organizational effectiveness through quality and strategy. Westport, Conn: Quorum Books; 1995.
69. ATLAS. ti qualitative data anaylsis URL: http://www.atlasti.com/index.html [accessed 2014-04-03] [WebCite Cache ID 6OYuNTpeB]
70. D'Amore JD, Sittig DF, Wright A, Iyengar MS, Ness RB. The promise of the CCD: Challenges and opportunity for quality improvement and population health. AMIA Annu Symp Proc 2011;2011:285-294 [FREE Full text] [Medline: 22195080]
71. Phillips N, Lawrence TB, Hardy C. Inter-organizational collaboration and the dynamics of institutional fields. Journal of Management Studies 2000 Jan;37(1):23-43. [doi: 10.1111/1467-6486.00171]
72. Contractor FJ, Lorange P. Trends in international collaborative agreements. In: Cooperative strategies in international business. Lexington, Mass: Lexington Books; 1988.

Abbreviations

CCD: continuity of care document
CHA: Community Health Alliance
EHR: electronic health record
HIE: health information exchange
Inova: Inova Health System
IOS: interorganizational system
IT: information technology
ONC: Office of the National Coordinator for Health Information Technology
PM: project manager
POC: proof of concept
STS: sociotechnical systems
VDH: Virginia Department of Health

Edited by G Eysenbach; submitted 06.04.14; peer-reviewed by B Tulu; comments to author 28.04.14; revised version received 09.06.14; accepted 10.07.14; published 15.08.14.

Please cite as:
Feldman SS, Schooley BL, Bhavsar GP
Health Information Exchange Implementation: Lessons Learned and Critical Success Factors From a Case Study
JMIR Med Inform 2014;2(2):e19
URL: http://medinform.jmir.org/2014/2/e19/
doi:10.2196/medinform.3455
PMID:25599991

Original Paper

Use of the Satisfaction With Amplification in Daily Life Questionnaire to Assess Patient Satisfaction Following Remote Hearing Aid Adjustments (Telefitting)

Silvio Pires Penteado[1*], PhD; Ricardo Ferreira Bento[1], PhD(Clin); Linamara Rizzo Battistella[2], PhD(Clin); Sara Manami Silva[1*], BA; Prasha Sooful[3*], MCP

[1]Medical School, Otorhinolaryngology Department, University of Sao Paulo, Sao Paulo, Brazil

[2]Medical School, Department of Forensic Medicine, Medical Ethics and Medicine and Social Work, University of Sao Paulo, Sao Paulo, Brazil

[3]Clinic, Audiology, Royal Darwin Hospital, Darwin, Australia

[*]these authors contributed equally

Corresponding Author:
Silvio Pires Penteado, PhD
Medical School
Otorhinolaryngology Department
University of Sao Paulo
Av Dr Eneas Carvalho de Aguiar 255
Suite 6167
Sao Paulo, 05403-000
Brazil
Phone: 55 11 30689855
Fax: 55 11 30689855
Email: penteadosp@gmail.com

Abstract

Background: Hearing loss can affect approximately 15% of the pediatric population and up to 40% of the adult population. The gold standard of treatment for hearing loss is amplification of hearing thresholds by means of a hearing aid instrument. A hearing aid is an electronic device equipped with a topology of only three major components of aggregate cost. The gold standard of hearing aid fittings is face-to-face appointments in hearing aid centers, clinics, or hospitals. Telefitting encompasses the programming and adjustments of hearing aid settings remotely. Fitting hearing aids remotely is a relatively simple procedure, using minimal computer hardware and Internet access.

Objective: This project aimed to examine the feasibility and outcomes of remote hearing aid adjustments (telefitting) by assessing patient satisfaction via the Portuguese version of the Satisfaction With Amplification in Daily Life (SADL) questionnaire.

Methods: The Brazilian Portuguese version of the SADL was used in this experimental research design. Participants were randomly selected through the Rehabilitation Clinical (Espaco Reouvir) of the Otorhinolaryngology Department Medical School University of Sao Paulo. Of the 8 participants in the study, 5 were female and 3 were male, with a mean age of 71.5 years. The design consisted of two face-to-face sessions performed within 15 working days of each other. The remote assistance took place 15 days later.

Results: The average scores from this study are above the mean scores from the original SADL normative data. These indicate a high level of satisfaction in participants who were fitted remotely.

Conclusions: The use of an evaluation questionnaire is a simple yet effective method to objectively assess the success of a remote fitting. Questionnaire outcomes can help hearing stakeholders improve the National Policy on Hearing Health Care in Brazil. The results of this project indicated that patient satisfaction levels of those fitted remotely were comparable to those fitted in the conventional manner, that is, face-to-face.

(JMIR Med Inform 2014;2(2):e18) doi:10.2196/medinform.2769

KEYWORDS

audiology; hearing aids; hearing loss; telemedicine; correction of hearing impairment; public policy; prosthesis fitting; telemedicine; questionnairies; quality improvement

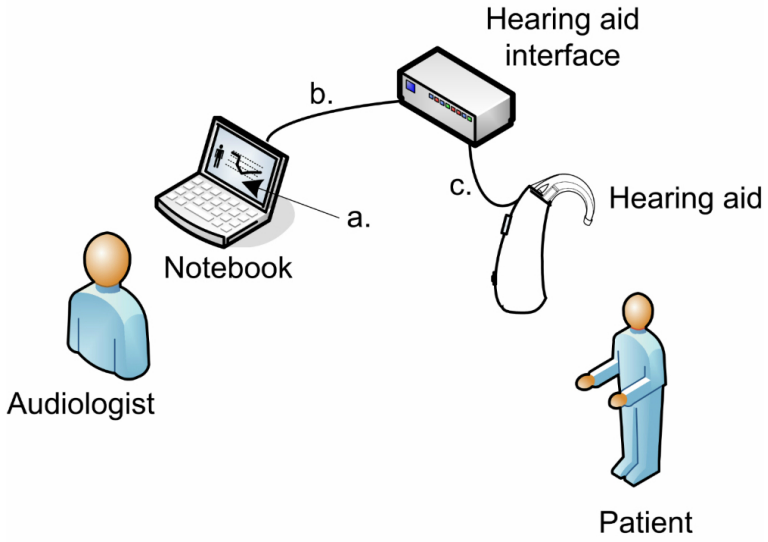
Introduction

The prevalence of hearing loss in Brazil has been identified as 50%, specifically, individuals with permanent hearing loss of more than 41 decibels hearing level (dB HL) in the municipality of Minas Gerais [1]. If this rate were extrapolated to the total Brazilian population, a contingent of more than 9 million Brazilians would be identified as having permanent hearing impairment. Since 2004, Brazil's public policy, National Hearing Health Care, has included the diagnosis, assessment, and rehabilitation of individuals with hearing loss. This policy culminated in the donation of hearing aids for Brazilian nationals seeking care at one of 139 centers accredited by the National Health System at various locations throughout the country. Many of the patients benefiting from this offer were from the adult and geriatric population. Patients were required to return to centers to receive necessary adjustments on their hearing aid settings and to receive information on care and usage of devices. The need for patients to return to the centers highlighted issues around transport to and from homes, cost of transport, as well as the need for carergivers to accompany patients.

The treatment of sensorineural hearing loss typically consists of the use of hearing aids (HA), and in cases of profound hearing loss, cochlear implants are used [2,3]. An HA is an electronic device that performs selective amplification, or amplifies signals at specific frequencies, using amplification strategies according to pathology and lifestyle of the hearing impaired [4]. The electronic architecture of a digital HA consists of a digital signal processor, microphone, and receiver [5].

Fitting an HA demands specific technical knowledge of the electroacoustic characteristics of the device in relation to the patient's hearing loss and, in Brazil, must be performed by an audiologist [6]. The fitting can be performed in a clinic or specialized hospital [7]. Digital HAs are fitted through an application known as a fitting program (Part a of Figure 1), which must be installed on a personal computer (PC). In this configuration (Figure 1), there must be a programming interface that connects the HA with the PC and the fitting application. A standardized programming interface commonly used is the HI-PRO device (GN ReSound A/S), which is connected to the PC via a universal serial bus (USB) cable (Part b of Figure 1). Finally, a programming cable is required to connect the HA to the interface (Part c of Figure 1).

Figure 1. Diagram of a basic adjustment with digital HAs.



After receiving an HA, it is essential to monitor the patient in order to understand changes that affect the auditory system [8] and thus support adjustments to HA settings while the patient acclimates. Subjective evaluation of the HA fitting can be accessed via patient questionnaires, such as the Satisfaction With Amplification in Daily Life (SADL), the International Outcome Inventory for Hearing Aids (IOI-HA), the Hearing Handicap Inventory for Adults, the Abbreviated Profile of Hearing Aid Benefit (APHAB), and the Hearing Inventory for Elderly (HHIE), among others.

Cox and Alexander [9] described the benefits of subjective self-assessment through patient questionnaires and scales. These reports yield valuable insights into the impact of impairment on everyday life and encourage the planning and execution of strategies to address the needs of the hearing impaired person. Additionally, self-reported outcome data can be used as a tool to validate the merit of a certain treatment program and can highlight areas of improvement. Aspects such as comfort, ability to hear with background noise, ease of hearing aid controls, ease of inserting and removing HAs, and so on, can be subjectively evaluated in the form of a satisfaction questionnaire. Objective assessment, however, requires the need for equipment and can be obtained by measuring the functional gain (the difference between the thresholds obtained with and without the HA in the free field mode), or by measuring the acoustic

gain by using a probe microphone, known as Real Ear Insertion Gain.

Telemedicine has evolved in many health care areas. With a greater coverage area and lower operating costs, it is fast becoming a suitable method for assessment, diagnosis, and rehabilitation of various conditions. In an early work with hearing rehabilitation through telemedicine, Wesendahl [10] reported that remote fittings allow for experienced professionals to be present "in remote areas without restriction of time and geographic location" and that "telemedicine offers an opportunity to increase the efficiency of audiological methods and decrease expenses simultaneously". Internationally, there have been many studies documenting the success of telemedicine by improving access to quality care [11,12]. In addition to hearing aids, cochlear implant remote mapping and adjustments [13-15] were highlighted by other studies [16,17]. Swanepoel et al [18] describes telemedicine: "although not the answer to all challenges related to global hearing loss, there is no alternative strategy that can offer the same positive impact on the current hearing loss burden in the near and foreseeable future." A literature review of the demonstrated benefits of telemedicine in diverse areas of medicine can be found in Table 1 [19-38].

Table 1. Some publications in diverse areas of telemedicine.

Authors	Area of science	Region	Results/ remarks
Chorbev & Mihajlov [19]	Several	Macedonia	Increased the population's access to health services, reducing costs, spread of knowledge to more distant centers
Jaakkola & Loula [20]	Public Policy	Finland	Decreased transport of patients and increased access to database of patients
Khaleel et al [21]	Several	United States	Blood pressure, heart rate, body temperature, blood glucose and ECG signals can be transmitted to remote centers in real time
Lavanya et al [22]	Dermatology	Singapore/United States	Dermatologists believed telemedicine to be beneficial when classroom visits were not possible or were troublesome
Penzel et al [23]	Management	Germany/ France/ Portugal	It was possible to establish a European network of Internet access among different clinics and other partners
Schreier et al [24]	Dermatology	Austria	Telemedicine using mobile phones equipped with camera enabled personalized therapy for psoriasis patients
Siddiqua & Awal [25]	Several	Bangladesh	Telemedicine was considered a way to improve the quality of health services, with improved access and lower costs
Shen et al [26]	Gynecology	United States	Preliminary studies have shown the effectiveness of developed systems, which improves the performance and diagnosis of breast diseases in remote areas
Stoian et al [27]	Disaster management	Romania	Telemedicine provided immediate results with greater chances than traditional methods
Sudhamony et al [28]	Oncology	India	Telemedicine offered great advantages in the practice of oncology as well as a decrease in the number of visits to emergency medical staff
Arriaga et al [29]	Neurology	United States	Telemedicine is a viable delivery model for neurotology care delivery
Audebert et al [30]	Cardiology	Germany/United States	Telemedicine recommended for the treatment of stroke
Bonato [31]	Rehabilitation	United States	The emergence of new sensors attached to the body capture the activity level of patients, helping the effectiveness of pharmacological interventions more efficiently and specifically
Capampangan et al [32]	Vascular	United States	The hit rate for decision conduit thrombosis in patients with acute stroke was broader with the use of telemedicine than with the use of telephone
Cardoso et al [33]	Cardiology	Brazil	Public efforts are key to implementing remote distance interventions for underserved populations in Brazil
Knobloch et al [34]	Reconstructive surgery	Germany	Using phone with HD camera delivers positive results in reconstructive surgery
Levine & Gorman [35]	Neurology	United States	Use of computer-based technology may be integrated with the neuroradiology, among others, to take care to distant areas
Mora et al [36]	Surgery	United States	Solution-based telemedicine can help in intermittent surgical services among patients and medical professionals
Mucic [37]	Psychiatry	Denmark	Patients preferred and recommended the use of telepsychiatry instead of psychiatry face-to-face with interpreters
Sacco et al [38]	Neurology	Italy	Patients with subarachnoid hemorrhage require the implementation of telemedicine in rural areas to minimize the high incidence of mortality

With regard to telemedicine nationally, a study conducted by the University of Sao Paulo described the effectiveness of video conferencing for transmitting video-laryngoscopic images [39]. While in another study with 73 subjects [40], researchers found that teleaudiometry proved to be an efficient method of hearing screening, with results close to a sweep audiometry, and that the use of teleaudiometry could help identify cases of hearing loss, especially in cases of patients with poor access to professionals.

A study of remote HA fittings [41] listed the benefits perceived by patients as those of ease and convenience in the fitting process, as well as less travel time. These factors may increase the outcomes of successful fittings and reduce social stigma. Bento and Penteado [42] theorized that remote HA fittings also benefit health care staff with regard to reduced travel time and costs.

The National Policy on Hearing Health Care in Brazil was established through Ordinance #2.073/04 GM (September 2004). Ordinance #402 (February 2010) of the Ministry of Health established the program nationwide (ie, Brazil Telehealth). Telemedicine in a structured format has aimed to quantify how to expand and strengthen strategies in family health. The Brazilian Federal Board of Audiology issued Resolution #366 (April 2009) that defined the lawful exercise of Telehealth in audiology with the use of information technology in order to "assist, promote education and conduct health research". According to official data, the government has become the largest purchaser of HAs in Brazil, as shown in Table 2.

Table 2. Investments in hearing health in Brazil (Ordinance #587 and #589).

Year	Total importation of HAs, units	Total purchases of HAs by the federal government, units	Percentage of purchases of HAs by the federal government, %
2005	169,575	113,983	67
2006	183,707	104,059	57
2007	214,310	134,194	57
2008	272,690	183,703	63
2009	280,578	184,646	66
2010	301,315	212,477	71
2011[a]	331,645	225,331	68
2012[a]	334,613	220,250	66
2013[a]	402,497	277,723	69

[a]Projection due to lack of official data.

Ordinance SAS/MS #58 specifies that HAs must be dispensed through centers accredited by the Unified Health System (SUS), where professionals must "perform diagnosis and rehabilitation of hearing loss in all age groups spanning neonates to geriatrics, and perform consulting ENT, neurological, pediatric audiological evaluation". Additionally, they must "ensure rehabilitation through clinical treatment in otolaryngology; selection, fitting and provision of an HA and speech therapy". The Ministry of Health has a list of 139 accredited centers to serve the population, which was 190,732,694 inhabitants divided into 8,514,876,599 km^2, according to the 2010 census.

Our research describes a pilot study conducted with 8 patients fitted remotely through telemedicine, using the Brazilian Portuguese version of the SADL as a tool for measuring subjective satisfaction, with the goal of improving hearing health policies in Brazil.

Methods

Approval

This research protocol was approved by the Ethics Committee for Analysis of Research Projects under #0293/11. The data collection was completed between June and October 2012.

Materials

The Brazilian Portuguese version of the SADL was used (see Multimedia Appendix 1); however, two SADL questions were discarded: #14: "Does the cost of your hearing aids seem reasonable to you?" and #15: "How pleased are you with the dependability (how often do they need repairs) of your hearing aids?" These deletions were justified because the participants did not buy their hearing aids and because it was not possible to evaluate the number of times the HA was sent out for repair, as only the initial and follow-up fitting appointments were conducted.

Additionally, Question 3 required a change as hearing aids were provided free of charge. The original question "Are you convinced that obtaining your hearing aids was in your best interests?" was changed to "Are you convinced that the received devices was the best option?" The documents used in this research are presented in Table 3.

Table 3. Study documents.

Document	Model	Version
Consent form	FMUSP	V 1.2
SADL questionnaire	Standard	Brazilian Portuguese
Terms and agreement of HA donation	TD	V 1.0

The Windows operating system was used for the data collection in this study because it has a large set of commercial applications available and is the largest PC platform used in Brazil. A broadband Internet connection was provided by the specialized unit (SU) and the remote unit (RU). The SU had a trained audiologist, who provided support and scientific training for the audiologist at the RU, thus acting as a facilitator.

The digital HAs donated and used this study were developed by researchers at the Medical School University of Sao Paulo, manufactured by Politec Saude (Multimedia Appendix 2). Only Behind-The-Ear (BTE) HAs were used in this study. A Portuguese version of the fitting application was supplied by the digital signal processor manufacturer ON Semiconductor. See Table 4 for the equipment we used and Table 5 for the applications used.

Table 4. Study equipment.

Description	Model	Manufacturer	Location
Notebook	Vostro 3500	Dell	SU
Notebook	Vostro 1510	Dell	RU
Hearing aid interface	HI-PRO	GN ReSound	RU
Router	78-0454ARB	GTS	SU
Router	ADSLCPE	ZTE	RU
Headphone	HT-301MV	Wasta	SU/RU
Web cam	1270	NA[a]	RU
Speakers	ND	FlexPc	RU

[a]Vostro 3500 Notebook has a built-in Web camera.

Table 5. Study applications and operating systems.

Name	Description	Version	Location
easyFIT	Hearing aids fitting	5.8.3.0	SU
TeamViewer	Remote access, VoIP[a]	7.0.14563	SU/RU
Medidor de velocidade de Internet	Internet speed meter on line	Full version	SU/RU
Operational system 32 bits	Windows 7	Professional	SU
Operational system 64 bits	Windows 7	Professional, Pack 3	RU

[a]VoIP: Voice over Internet Protocol. We chose VoIP by TeamViewer GmbH because it (1) had a free non-commercial version, (2) had compatible remote access, (3) allowed for message and file sharing, (3) allowed for recording sessions, (4) had a Portuguese version, (5) allowed adjustment of the microphone sensitivity, and (6) required minimal PC hardware requirements.

Participants

Participants were randomly selected through the Rehabilitation Clinical (Espaco Reouvir) of the Otorhinolaryngology Department Medical School University of Sao Paulo based on the following criteria: (1) male and female individuals aged between 18 and 90 years, (2) with either no obstruction of the external auditory canal or middle ear pathology, or an absence of any neurological or psychological impairment, (3) individuals with no prior HA experience, (4) bilateral sensorineural hearing loss of varying degrees (ie, mild, moderate, moderate-severe), (5) postlingual hearing loss, and (6) native Brazilians. Table 6 shows a summary of the participants in the study; five were female and three were male, with a mean age of 71.5 years.

It was important to include participants with no prior experience with HAs as long-term fitted subjects would have had difficulty answering Question 10 of the SADL, which relates to amplification. Furthermore, participants with no prior experience of amplification could make a judgment based solely on the amplification fitted in this study.

Table 6. Summary of participant data

Name (abbreviation)	Gender	Age (years)	Distance between home and remote unit (miles/km)
RRS	Female	83	3/4.8
APS	Male	85	14/22.5
FRO	Male	73	7/11.2
MCS	Female	56	9/14.5
GPS	Male	90	3/4.8
RNS	Female	48	8/12.8
NLSN	Female	59	12/19.3
MEC	Female	79	9/14.5

Design

We used an experimental research design. The following procedures were done face-to-face with the participants: (1) interview and otoscopy by otolaryngologist, (2) impedance and audiological measurements by audiologist, (3) agreement between patient and professional on the HA fitting, (4) earmold impressions, and (5) initial programming procedures with patient.

The following elements were remote (telefitting): (1) presence of an audiologist in RU, (2) presence of an audiologist in the SU (as the facilitator), (3) remote aid adjustments and changes to fitting data based on patient audiogram and subjective feedback, and (4) verification of patient satisfaction using the SADL questionnaire.

The SADL questionnaire was used as an interview schedule, that is, read aloud and completed by a trained interviewer for this purpose. The sessions were described as face-to-face (F) and remote assistance (R). There were two face-to-face sessions (F1 and F2) done within 15 working days of each other. The remote assistance (R) was 15 days after F2. This 15-day delay was justified so that patients would have adequate time and experience with the device and thus be able to respond accordingly to the SADL. At F1, patients agreed and signed the informed Consent Form (Chart 7). In the initial fitting sessions, the HAs were programmed through the easyFIT application, and the SU audiologist provided the necessary guidance to the patient. The face-to-face sessions followed the basic scheme described in Figure 1, while the remote service follows the basic scheme described in Figure 2.

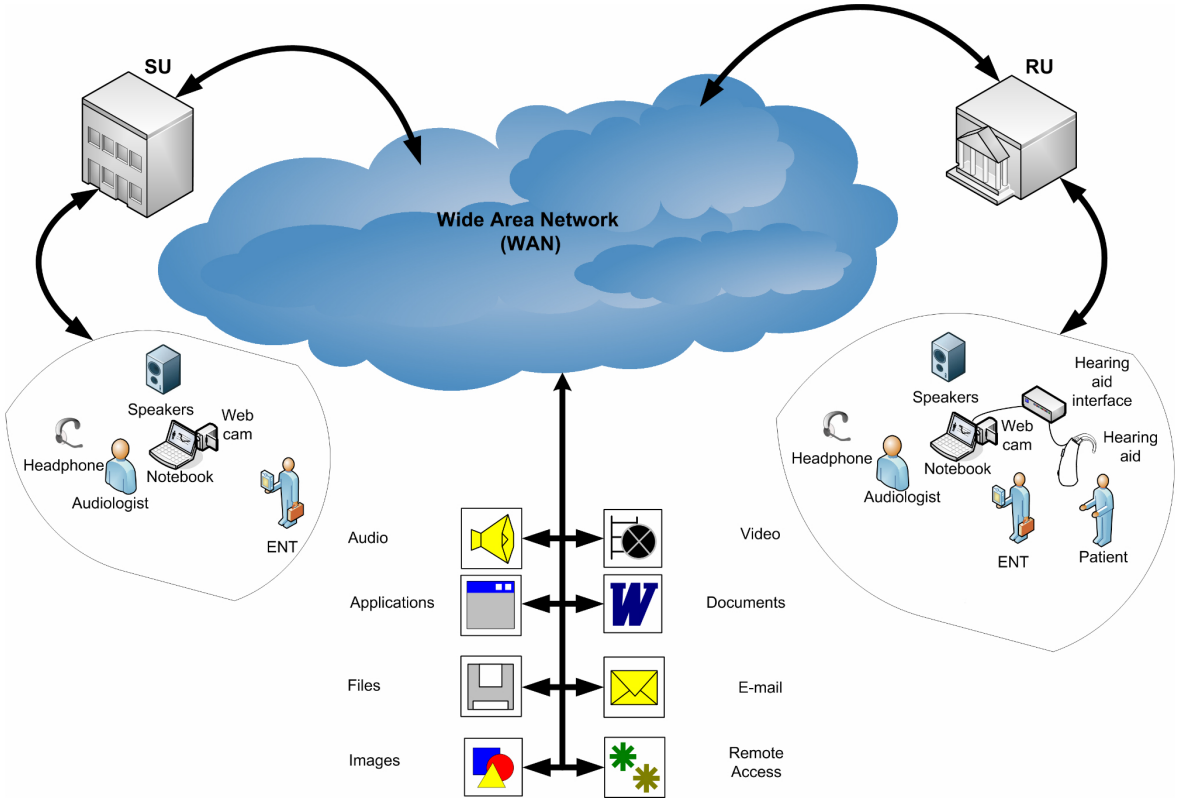
The SU had an audiologist trained to fit the HA Mini Retro C through easyFIT while the audiologist at RU had no specific training for the Mini Retro C, nor for easyFIT. Both units were supervised by an otologist, while 2 information technology professionals offered the technological support, one for each side.

For the remote (telefitting) sessions (Figure 2), all patients were based at the RU and accompanied by the RU audiologist, who explained that the SU audiologist would not be physically present but would interact with the patient through the Internet. The HAs were removed from the ears of patients to check for the presence of wax in the earmolds and then reinserted by the RU audiologist. The programming cables were inserted into their HAs, thereby allowing the programming interface to read the connected HAs, with identified model and serial numbers as well as access to all patient information (data personal, audiometry, and data sessions, etc).

The SU audiologist then remotely accessed the fitting application easyFIT and began procedures to check the HA settings. The patients were questioned subjectively about various aspects of their HA usage, which allowed the audiologist at SU to make adjustments remotely. In all cases, adjustments were required. Remotely, the SU audiologist recorded the new settings in the HAs and updated patient information in easyFIT. Additionally, situational/environmental advice was provided to the patient remotely before the SU audiologist ended their interaction with the patient.

Finally, the RU audiologist administered the SADL. The methodology of the questionnaire was explained, and each question was read aloud. This allowed for participants with poor literacy to fully understand all questions posed. Frequent pauses and repetitions were allowed for so that the patient had time to think about the answers. The session ended with the signing of Terms of Donation (TD - Chart 7) of HAs, again read aloud, so that patients with poor literacy were made aware of issues around the need to service HA quarterly from the date of signing the TD, as well as warranty periods, etc.

Figure 2. Basic scheme in telefitting session.



Results

In the initial telefitting session on September 18, 2012, 30 random measurements of Internet speed between the SU and RU were conducted (Table 7). Through the Datalogger feature for recording use of HAs, we found that over 50% (5/8) of patients used the HAs about 9 hours a day after being fitted, indicating good acclamation to the devices. Patient responses to the SADL questionnaire are presented in Table 8.

The results of the SADL from this study compared to four other studies are provided in Table 9 [43-45], which shows that the average scores from this study are above the mean scores from the original SADL normative data (Multimedia Appendix 3). These indicate a high level of satisfaction in participants who were fitted remotely.

Table 7. Internet speed in the specialized unit and the remote unit measured on September 18, 2012.

Parameters	RU			SU		
	Lowest	Highest	Average	Lowest	Highest	Average
Download[a] (kbps[b])	1521	4025	2842	9269	12,779	11,935
Upload[c] (kbps)	552	2554	1820	7496	11,160	9910
Ping[d] (ms[e])	34.4	95.6	58.5	4.2	11.5	7.3

[a]Download: speed (kbps) to download a particular file server.

[b]kbps: kilobyte per second = 1000 bits/ second; the digital signal transmission rate.

[c]Upload: speed (kbps) to load a particular file server.

[d]Ping: latency; the time (ms) necessary to test connectivity between information technology devices.

[e]ms: millisecond = 0.001 second.

Table 8. Summary of responses of patients to the SADL[a].

Patients	1	2	3	4	5	6	7	8	9	10	11	12	13
RRS	G	A	G	A	B	G	A	G	G	G	A	G	A
APS	G	A	G	A	G	G	A	G	G	G	G	G	A
FRO	G	A	F	A	G	G	A	F	G	F	F	G	F
MCS	G	F	G	B	E	G	A	G	G	G	D	G	A
GPS	G	A	G	G	G	G	A	G	G	E	E	G	A
RNS	G	A	G	A	G	G	A	G	G	E	G	G	A
NLSN	G	A	G	E	G	G	A	G	G	F	G	G	A
MEC	F	A	G	A	F	F	A	F	F	D	E	G	A

[a]A=not at all, B=a little, C=somewhat, D=medium, E=considerably, F=greatly, G=tremendously; see Multimedia Appendix 1.

Table 9. Results of our SADL compared with four other works (mean score and SD).

Factors	Cox and Alexander [9]	Danieli et al [43]	Mondelli et al [44][a]	Farias and Russo [45][a,b]	Our research (2012)
Positive effects	4.9 (1.3)	5.1 (1.3)	6.5 (0.5)	6.2 (0.8)	6.5 (0.4)
Negative features	3.6 (1.4)	4.5 (1.7)	6.3 (0.9)	6.2 (1.0)	6.2 (1.0)
Service and cost	4.7 (1.2)	5.5 (0.8)	4.7 (1.5)	6.7 (0.6)	7.0 (0.0)
Personal image	5.6 (1.1)	5.9 (0.9)	5.4 (1.6)	6.7 (0.4)	6.4 (0.7)
Average	4.7 (1.3)	5.2 (1.2)	5.7 (1.1)	6.4 (0.5)	6.5 (0.5)

[a]These authors presented a two-decimal precision of measurement, here rounded to only one decimal for purposes of comparison, according to National Center for Education Statistics NCES Standard: 5-3.

[b]These authors separated the results according to gender of patients. The results presented here are those of the larger group (males), although the gender difference is minimal.

Discussion

Principal Findings

Telefitting can help improve hearing health policies in Brazil by gradually expanding the current 139 accredited centers to centers closer to patients' homes as Basic Services Units, or health clinics, so the patient can receive adjustments to their HAs in their homes, with greater comfort and a greater chance of success in hearing rehabilitation. The participants in our study had a mean age of 71.5 years and had to travel an average of 8 miles (12.8 km) from home to RU. Table 9 reflects the patient satisfaction index close to one (6.5) to the maximum (7.0) with a low standard deviation (0.5), which indicates promising data on perceived benefits of patients fitted remotely. However, when compared with the study by Alexander and Cox [9], there is a considerable gap: 4.7 (1.3). This occurs due to the idiosyncratic difference between the two distinct audiences: the audience described by Cox and Alexander is based on individuals who had recently purchased HAs with their own resources in the United States, while the target populations in this study were SUS patients who received donated HAs. It is evident that patients with their own resources had different expectations than the patients in this pilot study.

Despite missing official data, it is possible to speculate that all the government-run health centers have at least one PC and probably an Internet account, which highlights the possibility of performing telefitting supported by an SU. There are 537 Basic Services Units (ambulatory level) throughout the city of Sao Paulo alone. Furthermore, it may not be necessary to have the hardware component HI-PRO, since it could be replaced by an application installed over the Internet. A universal programming cable for all HA manufacturers could be used instead of standard cables for a particular manufacturer. In general, most clinics have to operate with a large number of programming cables for various manufacturers.

The 2010 Brazilian Census reports that between 2005 and 2008, Internet access increased by 75.3% or 56 million users, due to various factors such as PCs and notebooks as well as high-speed Internet connections being more accessible and available at a lower cost. These factors, combined with other applications, can promote the use of telemedicine. Swanepoel et al [46] have highlighted new innovative means of bringing hearing health services to people through the benefit of telemedicine. Wasowski et al [47] concluded that the Nationwide Network of Teleaudiology with cochlear implants was a reliable platform for telefitting.

In this study, applications and user-level information technology were used, which although limited, allowed for the fitting of HAs in 8 patients. Nevertheless, if more advanced technology were used (eg, application-specific videoconferencing), the possibility of conducting real-time orthoscopic reviews would be increased. In addition, traffic-encrypted data over the Internet, access from other accredited PCs on the network, as well as integration with other applications, would ensure that management costs are kept to a minimum. There would also be

a reduced number of patient visits to the clinic and a database of valuable patient information to allow for remote HA adjustments.

In a more comprehensive telemedicine approach, it would be beneficial to include video training for the audiologist, as well as detailed training on the equipment and troubleshooting for complex fittings. An online tutor can assist in immediate cases (eg, when the audiologist at RU has doubts or is not familiar with fitting HAs), while full online courses on anatomy and physiology of hearing, interpreting audiometry, among others may be available.

Other questionnaires that could be used include the IOI-HA questionnaire, adapted to Brazilian Portuguese. It consists of seven questions, with a closed set of five different responses, and is thus easier to apply compared with SADL. The HHIE adapted to Brazilian Portuguese is structured into 25 questions, with a closed set of three possible answers on which Aiello et al [48] report that "further studies are needed to determine the convergent validity and construct validity of this instrument". Finally, the APHAB, which has not been adapted to Brazilian Portuguese, has 25 questions, each of which has 14 possible levels (each question includes two scenarios: with and without HA).

One can standardize the application of a satisfaction questionnaire, after the initial HA fitting and before the end of the warranty period. Thus, two questionnaires could record satisfaction levels at two significant times in the hearing rehabilitation process.

Strengths and Limitations

This study was one of the first of its kind with regard to adjusting HA settings via the Internet in Brazil. Furthermore, it creates a baseline for future research in this area of remote audiology and telefitting. However, our sample size of participants was small and within a limited geographical area. In addition, a limited number of applications were utilized.

Recommendations

In this study, certain applications and features were used to perform remote adjustments and fitting of HAs in 8 patients without injury. In the more comprehensive telemedicine approach, it would be necessary for the RU audiologist to have additional training and support. Further studies with a larger sample population should be conducted to explore the reproducibility of the results recorded here.

For implementation and public policy, we recommend the Windows platform be replaced by an open platform (eg, Linux, Ubuntu, Android) in order to reduce costs and promote the development of local solutions. The Internet services should ideally be linked to attributes such as stability, availability, speed, absence of risk, and confidentiality of patient data must be protected.

Furthermore, investigations conducted with the Brazilian Portuguese version of IOI-HA and APHAB questionnaires may be valuable. Although recording patient satisfaction through questionnaires provides valuable information about the use of HAs, the information derived is not entirely sufficient to assess the quality of overall hearing health. Bevilacqua et al [49] stated that "assessing the quality of health services is based in the infrastructure", that is, policies, facilities, and professionals. Silva and Formigli [50] generalized that in Brazil, the reorganization of health practices requires a definition of strategies for assessing changes of a broader management model.

A key step is to monitor satisfaction with technical issues by use of key performance indicators described by Kaplan and Norton [51], as a system of performance measurement and strategic management. As a result of understanding patients' difficulties, there can be continuous improvement in public health.

Conclusions

Remote HA adjustments (telefitting) have proved effective for these 8 patients, as indicated by their dynamic responses in SADL. Results were comparable to those of patients fitted in the conventional manner (ie, face-to-face fittings). Thus, the use of telefitting can be seen as an effective method to improve service delivery of hearing health in Brazil.

Acknowledgments

This work was supported by the Otorhinolaryngology Foundation.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Brazilian Portuguese version of the SADL.

[PDF File (Adobe PDF File), 289KB - medinform_v2i2e18_app1.pdf]

Multimedia Appendix 2

Mini Retro C Datasheet.

[JPG File, 2MB - medinform_v2i2e18_app2.jpg]

Multimedia Appendix 3

The SADL normative data.

[JPG File, 466KB - medinform_v2i2e18_app3.jpg]

References

1. Barak LR. Prevalencia de surdez incapacitante no municipio de Juiz de Fora, Minas Gerais. In: Published master's thesis. Sao Paulo, Brazil: University of Sao Paulo; May 14, 2011.
2. Bento RF, Miniti A, Marone SAM. In: University of Sao Paulo SP, Othrinolaryngology Foundation SP, editors. Tratado de otologia. Sao Paulo, Brazil: EDUSP; 1998.
3. Miniti A, Bento RF, Butugan O. Otorrinolaringologia Clínica e Cirúrgica. Rio de Janeiro, Brazil: Atheneu; 2000.
4. Carmen R. The consumer handbook on hearing loss and hearing aids. Sedona, Arizona: Auricle Ink; 2004.
5. Lybarger SF. A historical overview. In: Sandlin RE, editor. Handbook of hearing amplification. San Diego, CA: Singular; 1988.
6. Almeida K, Iório MCM, Dishtchekenian A. Próteses auditivas: uma revisão histórica. In: Almeida K, Iorio MCM, editors. Fundamentos teóricos & aplicações clínicas. Sao Paulo, Brazil: Lovise; 2003.
7. Sooful P, Dijk C, Avenant C. The maintenance and utilisation of government fitted hearing aids. Cent Eur J Med 2009 Feb 11;4(1):110-118. [doi: 10.2478/s11536-009-0014-9]
8. Momensohn-Santos TM. Testes para perdas auditivas funcionais. In: Momensohn-Santos TM, Russo ICP, editors. Pratica da Audiologia Clinica. Sao Paulo, Brazil: Cortez; 2005.
9. Cox RM, Alexander GC. Measuring Satisfaction with Amplification in Daily Life: the SADL scale. Ear Hear 1999 Aug;20(4):306-320. [Medline: 10466567]
10. Wesendahl T. Hearing aid fitting: application of telemedicine in audiology. Int Tinnitus J 2003;9(1):56-58. [Medline: 14763332]
11. Kokesh J, Ferguson AS, Patricoski C. Telehealth in Alaska: delivery of health care services from a specialist's perspective. Int J Circumpolar Health 2004 Dec;63(4):387-400. [Medline: 15709314]
12. Givens GD, Elangovan S. Internet application to tele-audiology--"nothin' but net". Am J Audiol 2003 Dec;12(2):59-65. [Medline: 14964319]
13. Gantz BJ, Turner C, Gfeller KE. Acoustic plus electric speech processing: preliminary results of a multicenter clinical trial of the Iowa/Nucleus Hybrid implant. Audiol Neurootol 2006;11 Suppl 1:63-68. [doi: 10.1159/000095616] [Medline: 17063013]
14. Davidson LS. Effects of stimulus level on the speech perception abilities of children using cochlear implants or digital hearing aids. Ear Hear 2006 Oct;27(5):493-507. [doi: 10.1097/01.aud.0000234635.48564.ce] [Medline: 16957500]
15. Eikelboom RH, Jayakody DMP, Swanepoel DW, Chang Samuel, Atlas MD. Validation of remote mapping of cochlear implants. J Telemed Telecare Online First 2014:1-7. [doi: 10.1177/1357633X14529234]
16. McElveen JT, Blackburn EL, Green JD, McLear PW, Thimsen DJ, Wilson BS. Remote programming of cochlear implants: a telecommunications model. Otol Neurotol 2010 Sep;31(7):1035-1040. [doi: 10.1097/MAO.0b013e3181d35d87] [Medline: 20147864]
17. Ramos A, Rodriguez C, Martinez-Beneyto P, Perez D, Gault A, Falcon JC, et al. Use of telemedicine in the remote programming of cochlear implants. Acta Otolaryngol 2009 May;129(5):533-540. [doi: 10.1080/00016480802294369] [Medline: 18649152]
18. Swanepoel de W, Clark JL, Koekemoer, Hall IIIJW, Krumm M, Ferrari DV, et al. Telehealth in audiology: the need and potential to reach underserved communities. Int J Audiol 2010 Mar;49(3):195-202. [doi: 10.3109/14992020903470783] [Medline: 20151929]
19. Chorbev I, Mihajlov M. Wireless telemedicine services as part of an integrated system for e-medicine. 2008 Presented at: 14th IEEE Mediterranean Electrotechnical Conference, MELECON; May 5-7, 2008; France p. 264.
20. Jaakkola H, Loula P. Managing a virtual hospital. 1996 Presented at: Conference on Engineering and Technology Management, IEMC; Aug. 18-20, 1996; Atlanta, Georgia.
21. Khaleel H, Al-Rizzo HM, Rucker DG. Wearable Yagi microstrip antenna for telemedicine applications. 2010 Presented at: Radio and Wireless Symposium (RWS), IEEE; Jan. 10-14, 2010; New Orleans, LA. [doi: 10.1109/RWS.2010.5434177]
22. Lavanya J, Goh KW, Leow YH, Chio MTW, Prabaharan K, Kim E, et al. Distributed Personal Health Information Management System for Dermatology at the Homes for Senior Citizens. In: Engineering in Medicine and Biology Society. 2006 Presented at: 28th Annual International Conference of the IEEE, EMBS; Aug. 30-Sept. 3, 2006; New York, NY.
23. Penzel T, Guilleminault C, Kesper K, Paiva T, Peter JH, Zulley J. The European neurological network. In: Engineering in Medicine and Biology Society. 1998 Presented at: 20th Annual International Conference of the IEEE, EMBS; 1998; Hong Kong.

24. Schreier G, Hayn D, Kastner P, Koller S, Salmhofer W, Hofmann-Wellenhof R. A mobile-phone based teledermatology system to support self-management of patients suffering from psoriasis. In: Engineering in Medicine and Biology Society. 2008 Presented at: 30th Annual International Conference of the IEEE, EMBS; Aug. 21-24, 2008; Vancouver, BC p. 5338-5341.
25. Siddiqua P, Awal MA. A portable telemedicine system in the context of rural Bangladesh. 2012 Presented at: International Conference on Informatics, Electronics & Vision (ICIEV); May 18-20, 2012; Dhaka, Bangladesh.
26. Shen Y, Pomeroy CA, Xi N, Methil-Sudhakaran N, Mukherjee R, Zhu D, et al. Supermedia interface for Internet-based tele-diagnostics of breast pathology. In: Biomedical Robotics and Biomechatronics (BioRob). 2006 Presented at: 1st IEEE/RAS-EMBS International Conference; Feb. 20-22, 2006; Pisa, Italy.
27. Stoian I, Golea A, Badea R, Moldovan A, Dancea O, Posteuca C. Using the cooperative robots concept in emergency and catastrophes medicine. 2006 Presented at: IEEE International Conference on Automation, Quality and Testing, Robotics; May 25-28, 2006; Cluj-Napoca, Romania.
28. Sudhamony S, Nandakumar K, Binu PJ, Issac Niwas S. Telemedicine and tele-health services for cancer-care delivery in India. IET Commun 2008;2(2):231. [doi: 10.1049/iet-com:20060701]
29. Arriaga MA, Nuss D, Scrantz K, Arriaga L, Montgomery E, St John P, et al. Telemedicine-assisted neurotology in post-Katrina Southeast Louisiana. Otol Neurotol 2010 Apr;31(3):524-527. [doi: 10.1097/MAO.0b013e3181cdd69d] [Medline: 20042903]
30. Audebert HJ, Schultes K, Tietz V, Heuschmann PU, Bogdahn U, Haberl RL, Telemedical Project for Integrative Stroke Care (TEMPiS). Long-term effects of specialized stroke care with telemedicine support in community hospitals on behalf of the Telemedical Project for Integrative Stroke Care (TEMPiS). Stroke 2009 Mar;40(3):902-908 [FREE Full text] [doi: 10.1161/STROKEAHA.108.529255] [Medline: 19023095]
31. Bonato P. Advances in wearable technology for rehabilitation. Stud Health Technol Inform 2009;145:145-159. [Medline: 19592792]
32. Capampangan DJ, Wellik KE, Bobrow BJ, Aguilar MI, Ingall TJ, Kiernan TE, et al. Telemedicine versus telephone for remote emergency stroke consultations: a critically appraised topic. Neurologist 2009 May;15(3):163-166. [doi: 10.1097/NRL.0b013e3181a4b79c] [Medline: 19430275]
33. Cardoso CS, Ribeiro AL, Castro RL, César CC, Caiaffa WT. Implementation of a cardiology care program in remote areas in Brazil: influence of governability. Rural Remote Health 2010;10(3):1472 [FREE Full text] [Medline: 20839899]
34. Knobloch K, Gohritz A, Vogt PM. Cell phone-based multimedia messaging service in reconstructive microsurgery: a novel telemedicine application. Plast Reconstr Surg 2009 Jun;123(6):220e-222e. [doi: 10.1097/PRS.0b013e3181a3f53b] [Medline: 19483557]
35. Levine SR, Gorman M. "Telestroke" : the application of telemedicine for stroke. Stroke 1999 Feb;30(2):464-469 [FREE Full text] [Medline: 9933289]
36. Mora F, Cone S, Rodas E, Merrell RC. Telemedicine and electronic health information for clinical continuity in a mobile surgery program. World J Surg 2006 Jun;30(6):1128-1134. [doi: 10.1007/s00268-005-0204-9] [Medline: 16736347]
37. Mucic D. Transcultural telepsychiatry and its impact on patient satisfaction. J Telemed Telecare 2010;16(5):237-242. [doi: 10.1258/jtt.2009.090811] [Medline: 20423935]
38. Sacco S, Totaro R, Toni D, Marini C, Cerone D, Carolei A. Incidence, case-fatalities and 10-year survival of subarachnoid hemorrhage in a population-based registry. Eur Neurol 2009;62(3):155-160. [doi: 10.1159/000226617] [Medline: 19571544]
39. Lazzarini CL. Análise da confiabilidade do telediagnóstico por imagens dinâmicas em laringologia. In: Published master's thesis. Sao Paulo, Brazil: University of Sao Paulo; 2004.
40. Campelo VE. Teleaudiometria: um método de baixo custo para triagem auditiva. In: Published master's thesis. Sao Paulo, Brazil: University of Sao Paulo, Brazil; 2009.
41. Penteado S, Ramos SL, Battistella LR, Marone SAM, Bento RF. Remote hearing aid fitting: Tele-audiology in the context of Brazilian Public Policy. Int Arch Otorhinolaryngol 2013 Dec 5;16(03):371-381. [doi: 10.7162/S1809-97772012000300012]
42. Bento RF, Penteado SP. Hearing rehabilitation through telemedicine to enhance public policies in Brazil. Einstein 2011;9(1):102-104.
43. Danieli F, Castiquini EAT, Zambonato TCF, Bevilacqua MC. Avaliação do nível de satisfação de usuários de aparelhos de amplificação sonora individuais dispensados pelo Sistema Único de Saúde. Revista da Sociedade Brasileira de Fonoaudiologia 2011;16(2):152-159.
44. Mondelli MF, Magalhães FF, Lauris JRP. Cultural adaptation of the SADL (satisfaction with amplification in daily life) questionaire for Brazilian Portuguese. Braz J Otorhinolaryngol 2011;77(5):563-572 [FREE Full text] [Medline: 22030962]
45. Farias RB, Russo JRP. Saúde auditiva: estudo do grau de satisfação de usuários de aparelho de amplificação sonora individual. Revista da Sociedade Brasileira de Fonoaudiologia 2010;15(1):26-31.
46. Biagio L, Swanepoel de W, Adeyemo A, Hall JW, Vinck B. Asynchronous video-otoscopy with a telehealth facilitator. Telemed J E Health 2013 Apr;19(4):252-258. [doi: 10.1089/tmj.2012.0161] [Medline: 23384332]
47. Wasowski A, Skarzynski H, Lorens A, Obrycka A, Walkowiak A, Skarzynski P, et al. The telefitting method used in the national network of teleaudiology: assessment of quality and cost effectiveness. Journal of Hearing Science 2012;2(2):81-85.

48. Aiello CP, Lima II, Ferrari DV. Validity and reliability of the hearing handicap inventory for adults. Braz J Otorhinolaryngol 2011;77(4):432-438 [FREE Full text] [Medline: 21860968]
49. Bevilacqua MC, Melo TM, Morettin M, Lopes AC. A avaliação de serviços em Audiologia: concepções e perspectivas. Revista da Sociedade Brasileira de Fonoaudiologia 2009;14(3):421-426.
50. Silva LMV, Formigli VLA. Avaliação em saúde: limites e perspectivas. Cad Saude Publica 1994;10(1):80-91.
51. Kaplan RS, Norton DP. The Strategy-Focused Organization: How Balanced Scorecard Companies Thrive in the New Business Environment. Boston, MA: Harvard Business School Publishing; 2000.

Abbreviations

APHAB: Abbreviated Profile of Hearing Aid Benefit
BTE: behind the ear
HA: hearing aid
HHIE: Hearing Inventory for Elderly
IOI-HA: International Outcome Inventory for Hearing Aids
PC: personal computer
RU: remote unit
SADL: Satisfaction with Amplification in Daily Life
SU: specialized unit
SUS: Unified Health System
USB: universal serial bus

Edited by G Eysenbach; submitted 14.06.13; peer-reviewed by W Glinkowski, A Holzinger; comments to author 13.10.13; revised version received 21.04.14; accepted 08.07.14; published 02.09.14.

Please cite as:
Penteado SP, Bento RF, Battistella LR, Silva SM, Sooful P
Use of the Satisfaction With Amplification in Daily Life Questionnaire to Assess Patient Satisfaction Following Remote Hearing Aid Adjustments (Telefitting)
JMIR Med Inform 2014;2(2):e18
URL: http://medinform.jmir.org/2014/2/e18/
doi:10.2196/medinform.2769
PMID:25599909

Original Paper

Using Business Intelligence to Analyze and Share Health System Infrastructure Data in a Rural Health Authority

Waqar Haque[1], MSc, PhD; Bonnie Urquhart[2], MSc; Emery Berg[3], BSc; Ramandeep Dhanoa[3], BTech

[1]University of Northern British Columbia, Department of Computer Science and School of Business, Prince George, BC, Canada

[2]Northern Health, Planning and Performance Improvement, Prince George, BC, Canada

[3]University of Northern British Columbia, Department of Computer Science, Prince George, BC, Canada

Corresponding Author:
Waqar Haque, MSc, PhD
University of Northern British Columbia
Department of Computer Science and School of Business
3333 University Way
Prince George, BC, V2N 4Z9
Canada
Phone: 1 250 960 6522
Fax: 1 250 964 4258
Email: waqar.haque@unbc.ca

Abstract

Background: Health care organizations gather large volumes of data, which has been traditionally stored in legacy formats making it difficult to analyze or use effectively. Though recent government-funded initiatives have improved the situation, the quality of most existing data is poor, suffers from inconsistencies, and lacks integrity. Generating reports from such data is generally not considered feasible due to extensive labor, lack of reliability, and time constraints. Advanced data analytics is one way of extracting useful information from such data.

Objective: The intent of this study was to propose how Business Intelligence (BI) techniques can be applied to health system infrastructure data in order to make this information more accessible and comprehensible for a broader group of people.

Methods: An integration process was developed to cleanse and integrate data from disparate sources into a data warehouse. An Online Analytical Processing (OLAP) cube was then built to allow slicing along multiple dimensions determined by various key performance indicators (KPIs), representing population and patient profiles, case mix groups, and healthy community indicators. The use of mapping tools, customized shape files, and embedded objects further augment the navigation. Finally, Web forms provide a mechanism for remote uploading of data and transparent processing of the cube. For privileged information, access controls were implemented.

Results: Data visualization has eliminated tedious analysis through legacy reports and provided a mechanism for optimally aligning resources with needs. Stakeholders are able to visualize KPIs on a main dashboard, slice-and-dice data, generate ad hoc reports, and quickly find the desired information. In addition, comparison, availability, and service level reports can also be generated on demand. All reports can be drilled down for navigation at a finer granularity.

Conclusions: We have demonstrated how BI techniques and tools can be used in the health care environment to make informed decisions with reference to resource allocation and enhancement of the quality of patient care. The data can be uploaded immediately upon collection, thus keeping reports current. The modular design can be expanded to add new datasets such as for smoking rates, teen pregnancies, human immunodeficiency virus (HIV) rates, immunization coverage, and vital statistical summaries.

(JMIR Med Inform 2014;2(2):e16) doi:10.2196/medinform.3590

KEYWORDS

business intelligence; health care systems; availability of health services; data visualization

Introduction

Health care extends beyond medicine in many ways. One of these is the ability to access health care services, particularly when one is located far from the core infrastructure. Access to relevant information in an intuitive form not only benefits the patient, but also assists the administration in identifying areas where resource allocation may have the highest impact. This ultimately leads to healthier communities and optimal use of health care funding. Fortunately, large volumes of information have been gathered over the years and this serves as a base for achieving the envisioned goals. Despite many recent government-funded initiatives, much of this information sits in legacy formats and the sheer volume of data makes it incomprehensible for any use other than the specific purpose for which each dataset was gathered. In addition, the data is of poor quality, suffers from inconsistencies, and lacks integrity. Despite having the data, health care providers and supporting staff are faced with the challenge of determining the type and location of resources accessible to them and their patients. In order to locate this information, an extensive search through numerous Microsoft Excel workbooks, databases, and statistical websites is quite common. Even then, it can be extremely tedious to find the needed information from these sources because they generally differ in purpose and tend to be inconsistent with each other.

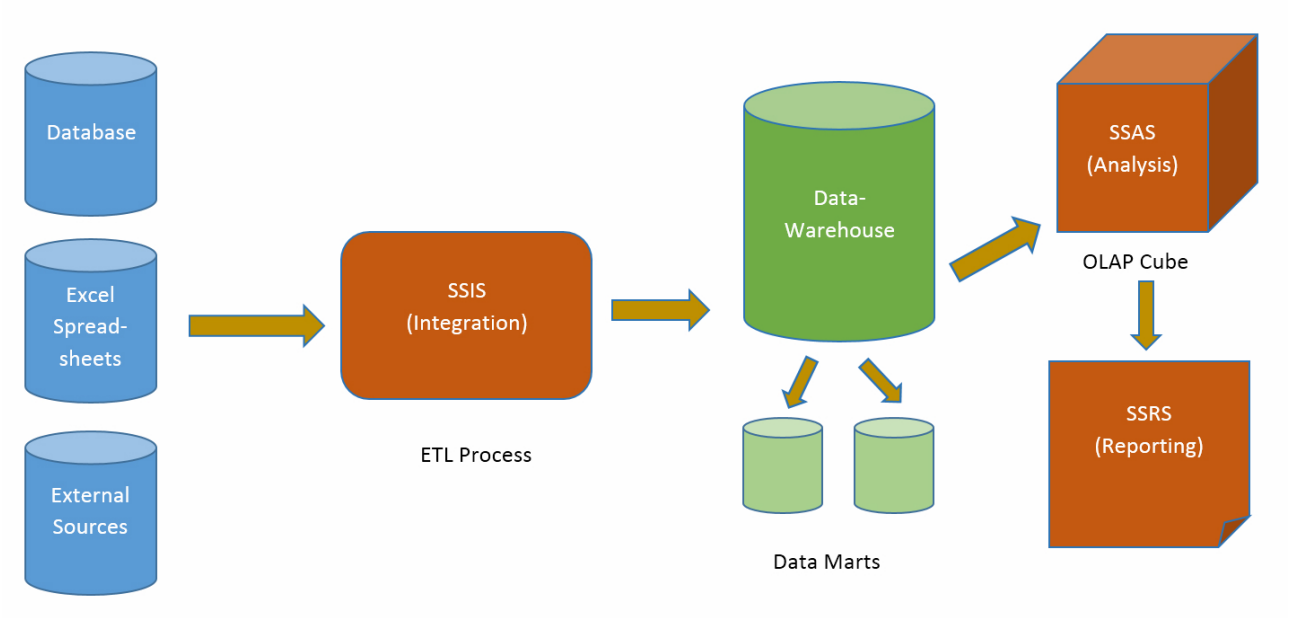
Traditionally, Business Intelligence (BI) has been used to analyze business information such as marketing and/or financial reporting data. In this paper, we propose how BI techniques can be applied to health care infrastructure data in order to make this information more accessible and comprehensible for a broader group of people. Our envisioned goal has been achieved by first consolidating the sources into a singular entity, second providing interactive access and control of the underlying data, and finally visually representing the data through reports. By applying these techniques, the resulting information can be accessed through dashboards, which provide a quick overview of the key performance indicators (KPIs) and allow navigation to underlying reports of finer granularity. Thus, instead of sifting through massive spreadsheets for the desired information, one can now access a centralized system that renders reports in a matter of seconds. The system also extends to other tools such as Web forms for updating data by designated staff without the need for going through complex IT protocols. The underlying data represents geography and services for the entire region covered by Northern Health (NH), that is, a population of approximately 300,000, land mass of roughly 600,000 km^2, and the breadth of services from health prevention and promotion through to acute care services. Northern Health is located in British Columbia, Canada, and is one of the seven health authorities in the province responsible for delivery of publicly funded health services. Five of the health authorities are based on geography, one is responsible for province-wide tertiary services, and one is responsible for First Nations health services. The Canadian health care system is publicly funded for the most part, with funding from both the federal government and the provincial or territorial governments.

Methods

Data Integration, Analysis, and Reporting

Business intelligence tools and techniques are an effective way to integrate and analyze large data repositories. However, the integration process becomes challenging when the data is not collected with analytics in mind. In our solution, we used Microsoft SQL Server's BI tool stack [1] and Web development framework, ASP.NET [2], to make the data more accessible and reduce the time that data analysts spend searching through large collections of sources. We also merged the disparate data sources to eliminate data conflicts and create a singular source for reporting. The resulting data warehouse became the central source for all analysis and reporting (Figure 1). An extract-transform-load (ETL) [3] process was used to populate the data warehouse. During the extract phase, connections were created to various data sources and the required information was pulled into temporary storage. In the transform phase, the format of stored data was made consistent with metadata prior to loading into the data warehouse. The SQL Server Integration Services (SSIS) component in Microsoft's BI tool stack was used to accomplish this integration using an ETL process. SSIS provides the ability to fetch data from disparate sources and apply different transformations on the data, for example, convert from one data type to another, alter data in the sorted order, etc. An Online Analytical Processing (OLAP) cube was then created using the Analysis Services. This cube is an n-dimensional structure, which can be used to reveal more complex details at various levels of granularity through predesigned and ad hoc queries. The cube consists of several dimensions and fact tables [3]. Using the cube structure, reports are created and rendered through Microsoft's reporting services [1]. SQL Server Reporting Services provides a rich set of data visualization features such as charts, tables, matrices, gauges, maps, and tooltips. A dashboard gives a high-level overview of the KPIs and acts as a central navigation hub to other reports. Mapping allows the user to see the information based on regions and provides visual representation of distances between locations. Web forms are used to allow users to remotely update information with automatic consistency checks. Normally, updates to a database require knowledge of the underlying structure and the associated query language. By providing a Web form, these queries are created automatically and the database structure is represented visually for easy understanding.

Figure 1. Business intelligence modeling overview.



Related Work

Historically, the health care field has been slow to adopt new computer technologies; this has been largely due to hardware limitations, insufficient computer literacy, mechanical user interfaces, and privacy concerns. The first two causes have been mostly overcome due to the penetration of computers in daily lives and the technological advancements in computer hardware, but many applications are still mechanical in nature and are not intuitive to the user [4]. The Minnesota Health Association developed a pilot program to combine clinical information with administrative data, which faced many challenges such as the expertise of those involved and communication issues resulting from distributed data sources [5]. BI tools and techniques have been used to provide insight into ambulatory care sensitive conditions within Northern Health by analyzing data and identifying areas that need attention [6]. These techniques have also been used successfully to improve the management of large quantities of medical information [7]. Historical information and comparisons with the United States' primary care system has shown that providing improved access to primary care reduces the cost of health care and enhances the care provided to patients [8]. A comparison survey observed that with the increased access to health care in Canada, the general health of the public was superior to that in the United States [9]. Another study showed that though Canada has a relatively lower cost of health care, the wait times negatively affect the perceived availability of care [10]. Additionally, disparities in health care and its access have been shown to be negatively related to lower income, education, and race both due to perception and access [11]. There has been little or no significant evidence of work that incorporates the concept of BI in analysis of data related to asset mapping or services availability.

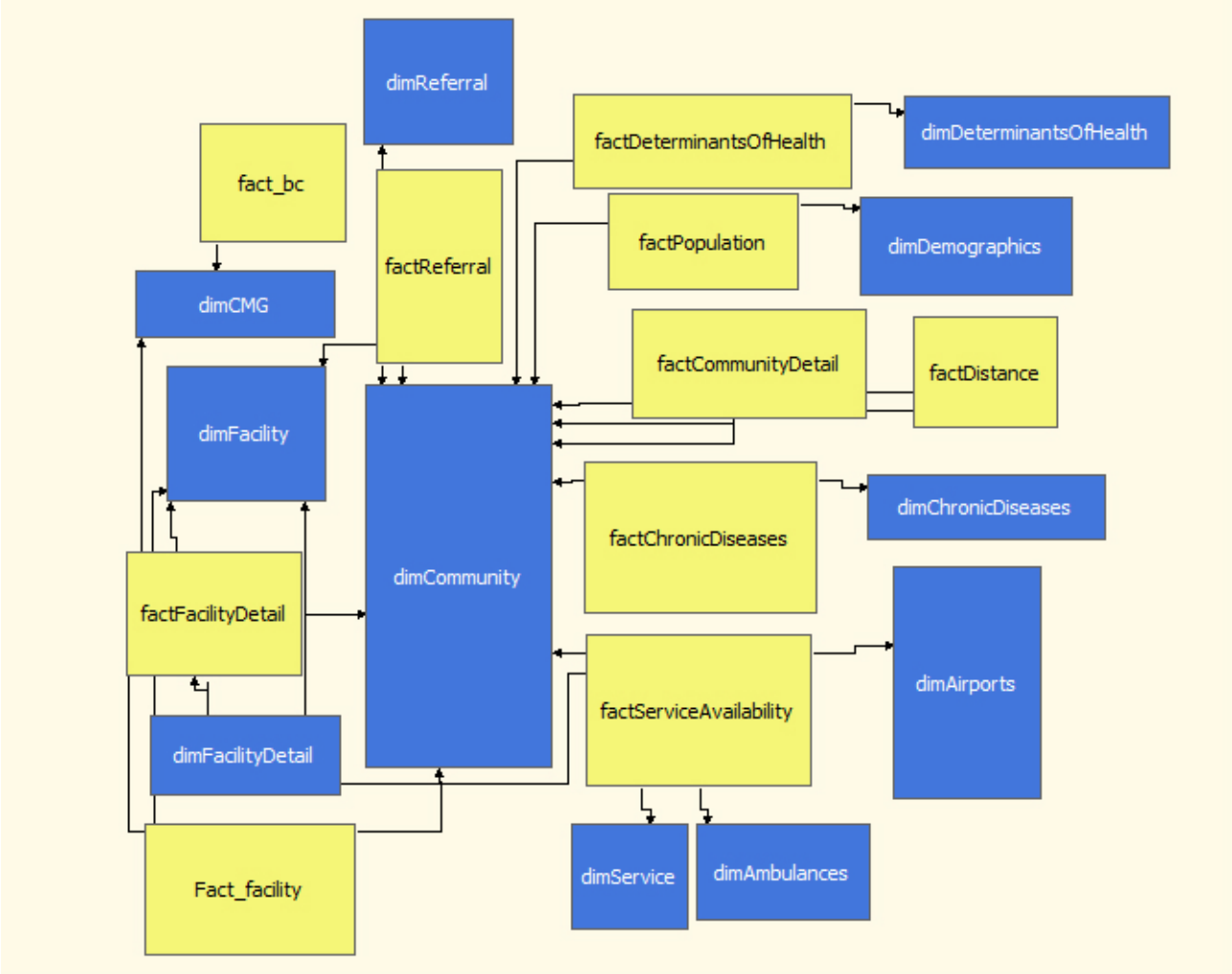
Data Challenges and Cube Design

The underlying data was collected over several years for varying purposes including generation of community health reports. The complexity of the underlying data posed several challenges in the integration phase. The first and foremost challenge was the sheer number of workbooks (a workbook can have many Microsoft Excel spreadsheet files), which have been the primary source of information for the data analysts for several years. An initial screening eliminated irrelevant data, but even after this exercise a very large number of workbooks remained. Most of these contained several sheets that were created for a variety of (sometimes unrelated) purposes, which meant the data did not always match in content or level of granularity. Even the repeated numbers sometimes differed across the workbooks. The data was available at various levels of hierarchy making aggregations unpredictable. Similarly, different naming schemes were used for locations without specifying any clear relationship(s) among them. To deal with this, fuzzy lookups [3] were used by specifying a threshold to match names that are similar enough but not identical. For locations that failed to match, a manual mapping table was created and the names were corrected at the database level through SQL queries.

A relational model was developed to create a singular source of information. This database consisted of 24 relations, which were populated by three dump sheets via Web forms (described later). An integration package was built to cleanse, combine, and group the data based on its purpose and granularity. When there were conflicts due to repeated information, the selection was based on conformity with other sources and the age of data. In rare cases, informed calculations were performed to correctly reflect missing values. The next step was to create an analysis cube using this database. Normally such cubes use star schema with a single fact table and multiple dimensions [3]. While this structure gives superior performance due to a reduced need for joins, it requires all information to exist at the same level of granularity. This was impractical in our case because of the need for added rows and empty cells if the data were to be restructured. Thus, in our somewhat unusual design, 10 fact tables and 11 dimensions were used (Figure 2), primarily for performance and granularity reasons.

Another challenge was to allow seamless update of data by analysts and staff unfamiliar with the underlying schemas and not trained to write sophisticated queries. Besides having a capability for bulk loading of large volumes of data, there was also a need for the ability to update individual rows without affecting the integrity of the database. To provide this functionality, a Web form was created in ASP.NET [2]. An intuitive combination of tabs, groupings, and dropdown lists allow data entry into individual cells of the selected table. The entered values were checked against metadata before updates were committed. Another mechanism to prevent inconsistencies was the use of dropdown lists when names were referenced. For bulk loading, two dump sheets reflecting the database structure were created. These sheets allow data to be compiled or manually entered and uploaded to the Web form. An integration process then triggers to transparently upload the data and reprocess the cube. The integration is fast, stable, and reliable due to instant validation and simplified logic. The data was also validated by sending reports to administrators of relevant health service delivery areas and by checking against existing manually generated reports.

Figure 2. The Online Analytical Processing (OLAP) cube.



Results

Reporting

For interactive access to information and data visualization, Microsoft SQL Reporting Services [1] were used to generate dynamic reports. In addition to conventional charts and graphs, access to advanced features like mapping, navigational controls, and parameterization of reports is also provided. The information contained in these reports can be updated through the Web form, which automatically reprocesses the cube and immediately reflects the changes.

Main Dashboard

The main dashboard provides an overview of the KPIs and includes navigation controls including a toggle control to switch between demographic information, patient profiles, and case mix groups (CMG). The demographic information displays information relevant to the population status including factors such as wealth, education level, origin [12], and dependency rates (Figure 3). These metrics assist in identifying potential areas of concern and any needed level of support and services. The patient profile gives an overview of the health-related metrics within the selected region by showing information such as births, commonality of chronic conditions, and vaccine preventable diseases (Figure 4). The CMG profile ranks the top

20 reasons for hospitalization in the selected geographic region as compared with the entire province (Figure 5).

This information is available at all levels of hierarchy with the granularity becoming finer from Northern Health Authority (NHA) to Health Service Delivery Areas (HSDA) to Local Health Authorities (LHA) to Communities. This hierarchy can be selected from the maps, which in turn generates the parameters for necessary filtration of information. Tabbed controls allow switching to other reports while maintaining the current level of hierarchy. These tabs allow access to subgroups such as availability of services, comparisons of selected regions, service levels, and direct access to community profiles. To improve navigational performance, the header is embedded with parameters to track the current tab select, type of report (drill-down), and the current level. This allows transparent passing of parameters to determine which report or level to load next based on the direction of navigation.

Figure 3. Main dashboard: Population profile.

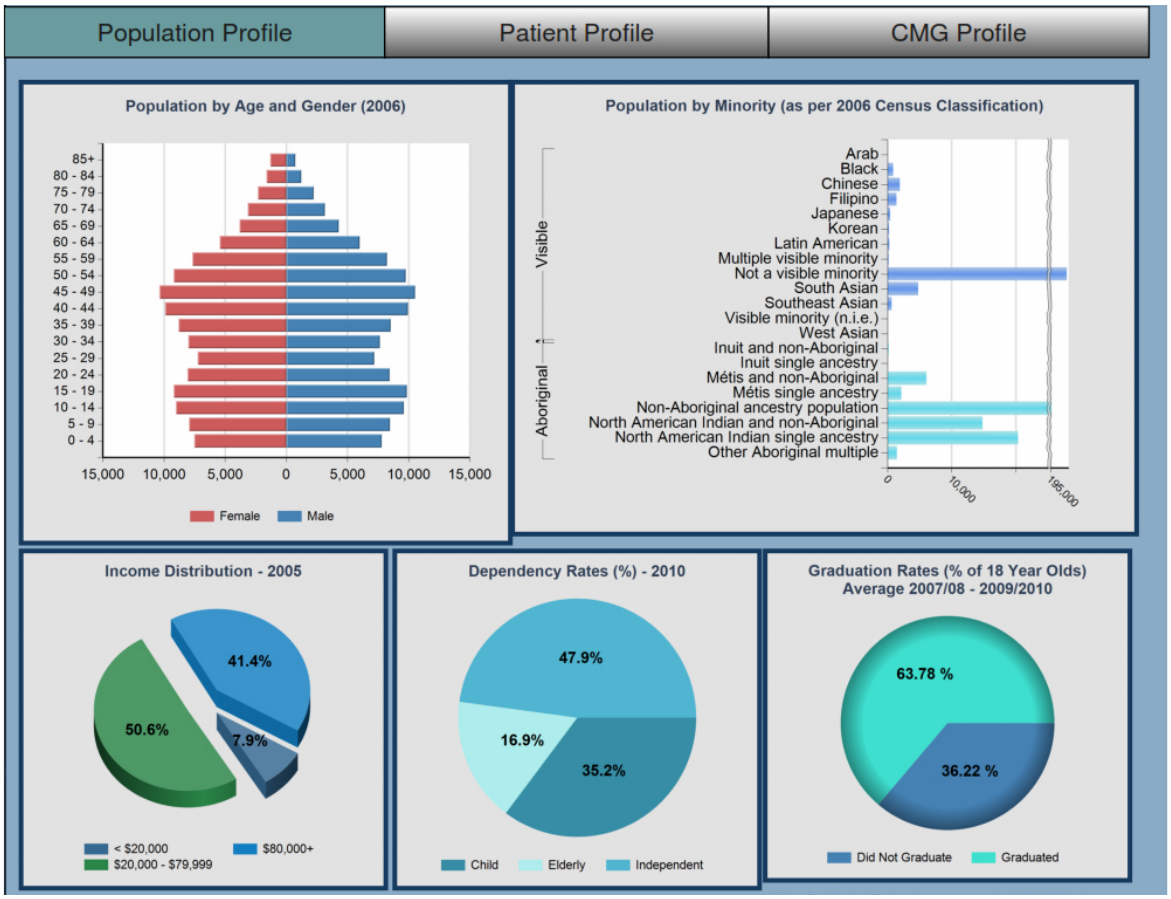


Figure 4. Main dashboard: Patient profile.

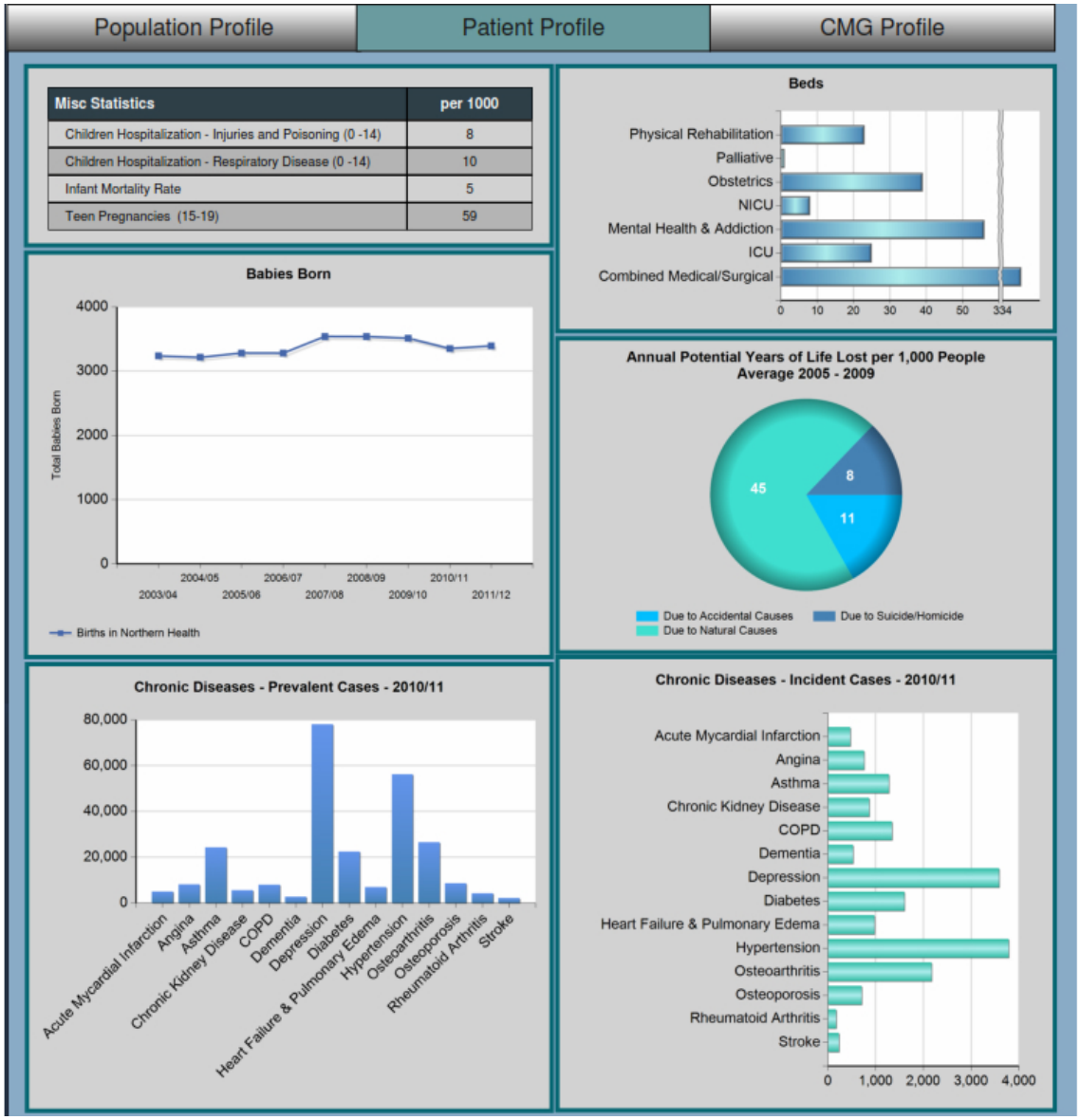
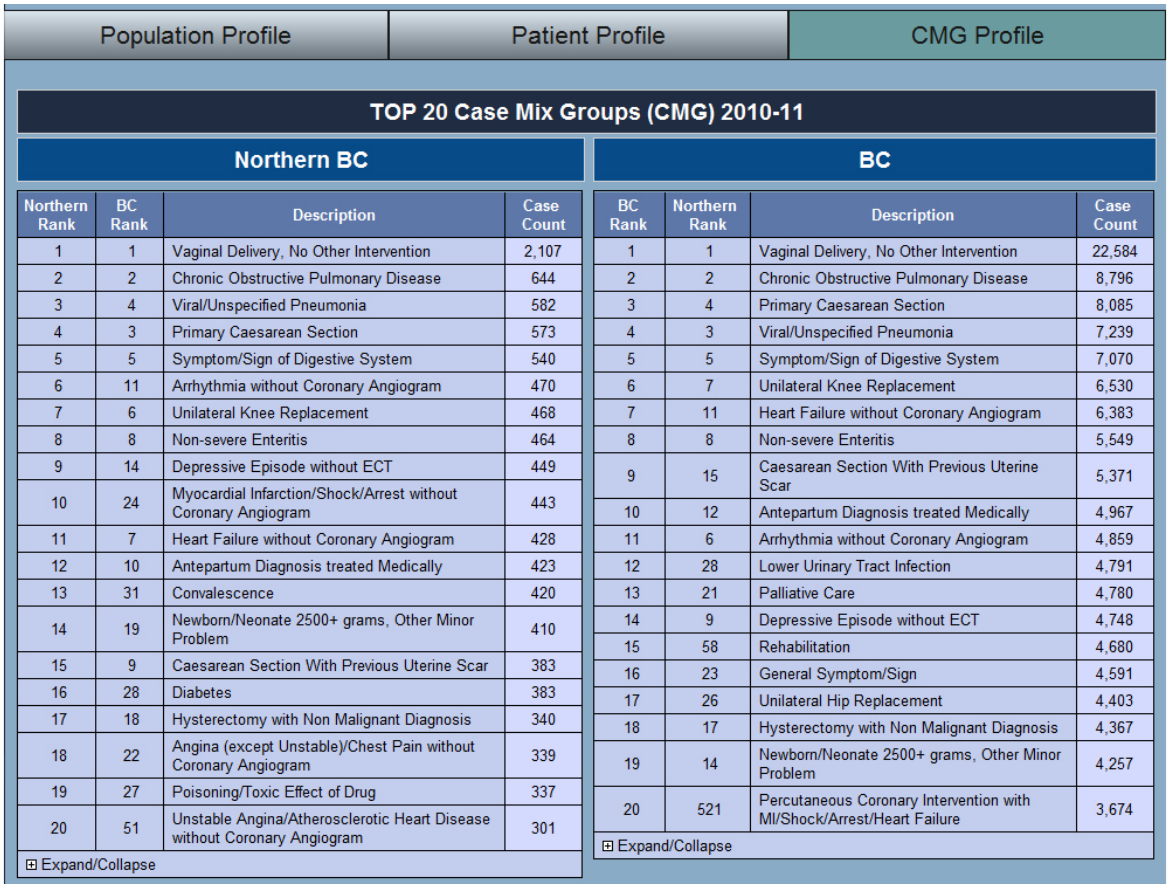


Figure 5. Main dashboard: Case Mix Groups (CMG) profile.



Population Profile | Patient Profile | CMG Profile

TOP 20 Case Mix Groups (CMG) 2010-11

Northern BC

Northern Rank	BC Rank	Description	Case Count
1	1	Vaginal Delivery, No Other Intervention	2,107
2	2	Chronic Obstructive Pulmonary Disease	644
3	4	Viral/Unspecified Pneumonia	582
4	3	Primary Caesarean Section	573
5	5	Symptom/Sign of Digestive System	540
6	11	Arrhythmia without Coronary Angiogram	470
7	6	Unilateral Knee Replacement	468
8	8	Non-severe Enteritis	464
9	14	Depressive Episode without ECT	449
10	24	Myocardial Infarction/Shock/Arrest without Coronary Angiogram	443
11	7	Heart Failure without Coronary Angiogram	428
12	10	Antepartum Diagnosis treated Medically	423
13	31	Convalescence	420
14	19	Newborn/Neonate 2500+ grams, Other Minor Problem	410
15	9	Caesarean Section With Previous Uterine Scar	383
16	28	Diabetes	383
17	18	Hysterectomy with Non Malignant Diagnosis	340
18	22	Angina (except Unstable)/Chest Pain without Coronary Angiogram	339
19	27	Poisoning/Toxic Effect of Drug	337
20	51	Unstable Angina/Atherosclerotic Heart Disease without Coronary Angiogram	301

⊞ Expand/Collapse

BC

BC Rank	Northern Rank	Description	Case Count
1	1	Vaginal Delivery, No Other Intervention	22,584
2	2	Chronic Obstructive Pulmonary Disease	8,796
3	4	Primary Caesarean Section	8,085
4	3	Viral/Unspecified Pneumonia	7,239
5	5	Symptom/Sign of Digestive System	7,070
6	7	Unilateral Knee Replacement	6,530
7	11	Heart Failure without Coronary Angiogram	6,383
8	8	Non-severe Enteritis	5,549
9	15	Caesarean Section With Previous Uterine Scar	5,371
10	12	Antepartum Diagnosis treated Medically	4,967
11	6	Arrhythmia without Coronary Angiogram	4,859
12	28	Lower Urinary Tract Infection	4,791
13	21	Palliative Care	4,780
14	9	Depressive Episode without ECT	4,748
15	58	Rehabilitation	4,680
16	23	General Symptom/Sign	4,591
17	26	Unilateral Hip Replacement	4,403
18	17	Hysterectomy with Non Malignant Diagnosis	4,367
19	14	Newborn/Neonate 2500+ grams, Other Minor Problem	4,257
20	521	Percutaneous Coronary Intervention with MI/Shock/Arrest/Heart Failure	3,674

⊞ Expand/Collapse

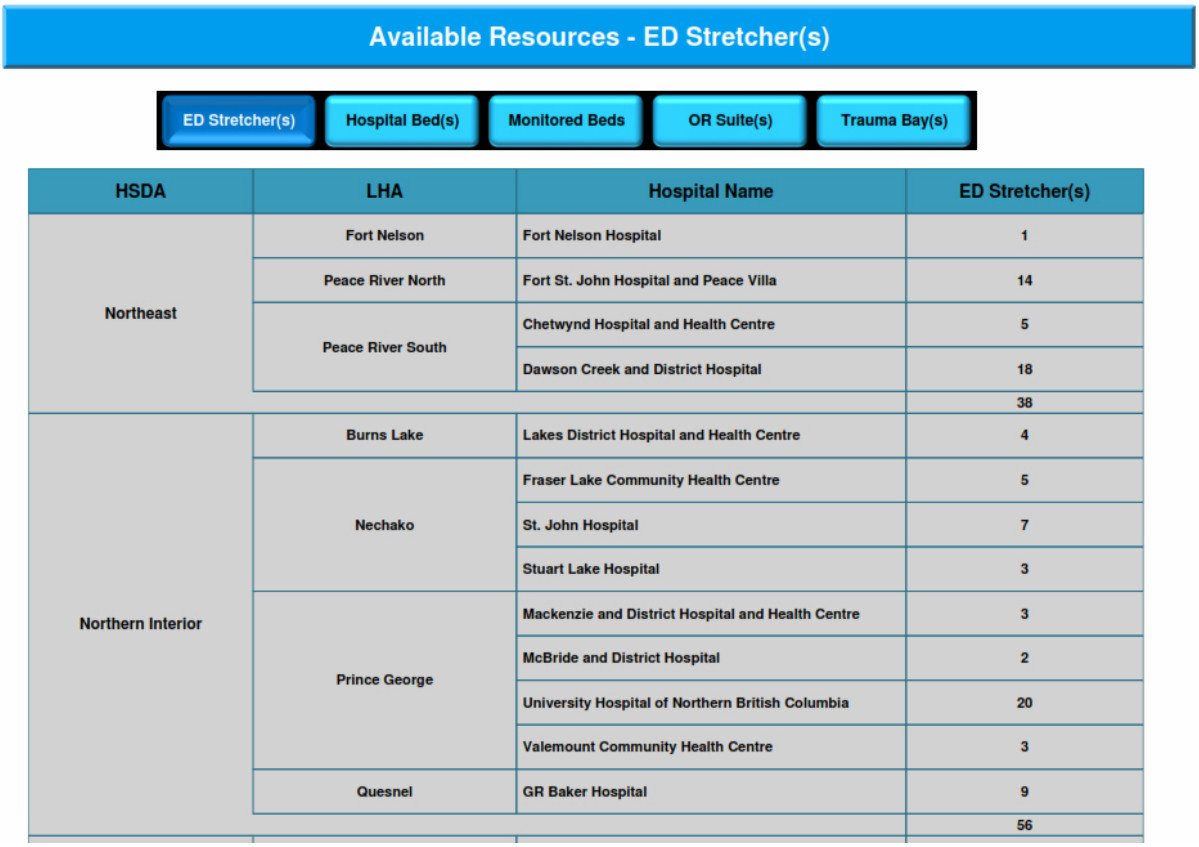
Critical Care Dashboard

The Critical Care dashboard is an extension of the main dashboard; however, as the combined project progressed it became necessary for the two to split. The reasons for the split included level of granularity desired, validation of data by different groups, and the inward facing (to institution) nature of the Critical Care dashboard as opposed to the outward facing (to public) for the Services Availability dashboard. This dashboard shows the availability of resources for the hospitals and health centers. The ideas and approaches used in the main dashboard have been carried over to the Critical Care dashboard. It opens up to an overview of the NHA area, having drill-down capabilities to HSDA and then to LHA level through a map (Figure 6). Main metrics of the dashboard are Available resources and staff, Number of hospitals with staffing needs, Accepting and transferring patients, Staff credentials and connections, Bed-line transfer use, Airway support, Ventilator capacity, Ventilating patients, Respirators, Ambulances, First Nation Communities, etc. Since all the facilities do not input data, the number of facilities reporting data is described at the top right side of the page. All metrics on this page have drill-down capabilities for easier analysis (Figure 7).

Figure 6. Critical Care dashboard.

	24 hours	Limited hours	Onsite
CT	-	5	6
Lab	1	23	24
MRI	-	1	1
Radiology	-	21	23

Figure 7. Available resources - drilldown report.

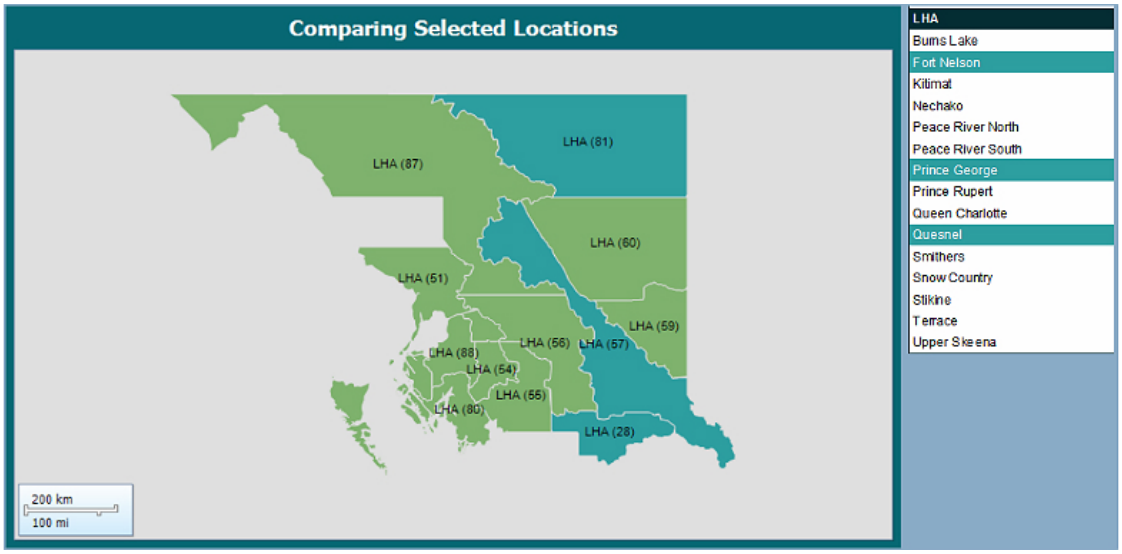


HSDA	LHA	Hospital Name	ED Stretcher(s)
Northeast	Fort Nelson	Fort Nelson Hospital	1
	Peace River North	Fort St. John Hospital and Peace Villa	14
	Peace River South	Chetwynd Hospital and Health Centre	5
		Dawson Creek and District Hospital	18
			38
Northern Interior	Burns Lake	Lakes District Hospital and Health Centre	4
	Nechako	Fraser Lake Community Health Centre	5
		St. John Hospital	7
		Stuart Lake Hospital	3
	Prince George	Mackenzie and District Hospital and Health Centre	3
		McBride and District Hospital	2
		University Hospital of Northern British Columbia	20
		Valemount Community Health Centre	3
	Quesnel	GR Baker Hospital	9
			56

Mapping Functionality

The mapping features allow location-based visualization and navigation. A challenge, however, was the lack of availability of full range of maps. While default maps are provided by the tools, no maps of British Columbia (BC) were included; this resulted in the need for a shape file to store geographic information such as the shape of regions, locations of communities, or other geographic features. The shape files of BC were obtained from [13] and modified to better fit our needs. These modifications were done through an open source geographic information system (GIS) application, QuantumGIS [14]. Using this application, the shape file of BC was restricted to the area covered by NH; another shape file was created for storing the locations of communities within the region. To address stability issues created by the large size of the single shape file, steps were taken to limit the information contained therein, primarily by removing information that was available elsewhere. Maps were also used for controls in the comparison report and to visualize availability of services in the community or proximity. An example of these controls can be seen in Figure 8 where the map has been used to select LHAs for comparisons.

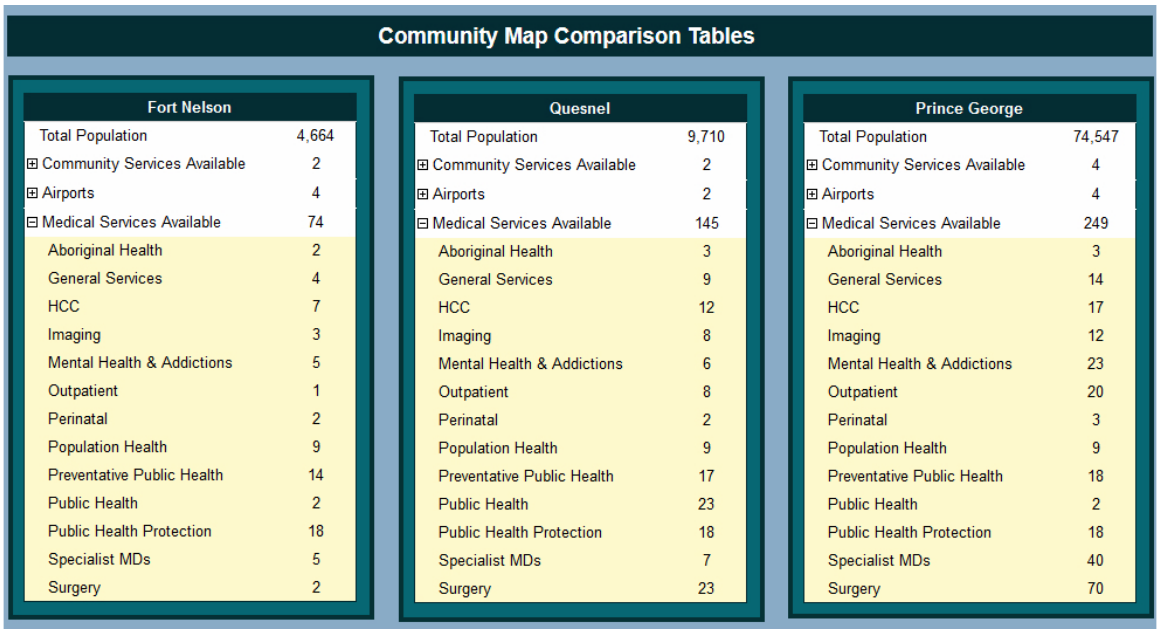
Figure 8. Comparison map.



Comparison

Comparisons between differing regions provide further insight into the state of health care within a region and potential causes for disparities [9]. The regions of interest can be selected by simply clicking on the map. The comparison can be performed at all levels and allows for comparing up to three regions at a time (Figure 8). When a new region is selected for comparison, the three most recent selections are maintained. Metrics such as population, facilities, community services (airports, ambulances, etc), and medical services are displayed for each selected region (Figure 9).

Figure 9. Community comparison tables.



Service Availability

Examining the demographic information, patient details, and availability of services has been shown to be effective in identifying possible needs of a region and improving the care of patients [8,10,11,15]. This information has been provided using color-coded markers on the map. For readability purposes, we show up to four services in circles split into quarters (Figure 10). The services are selected from a categorized list; additional information about the locations is displayed through tooltip display when hovering over the circles. Each time a service is selected, the map is updated to show the communities that have facilities offering the selected services. All services offered in NH, whether or not those are available locally, can also be seen in the availability report at the community level. This expandable list shows the proximity of where missing services can be found and the distance/travel time from the current community (Figure 11).

Figure 10. Available map.

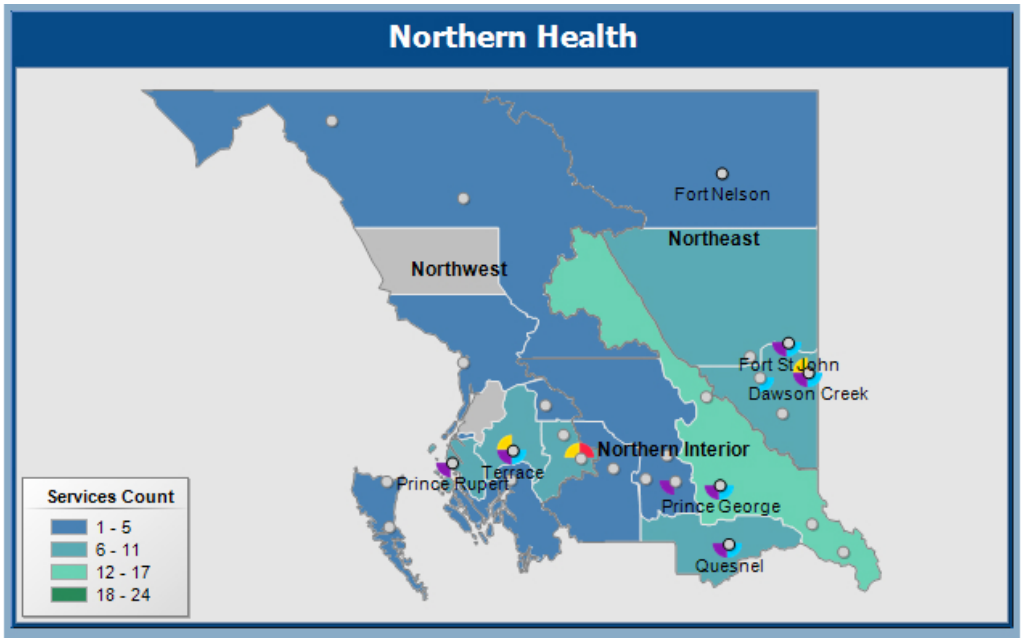


Figure 11. Community profile: Service availability.

Community Level Reports

The Community Profile (Figure 12) summarizes demographic information, chronic conditions, hospitalization rates, available services, and other health-related metrics. It can be browsed from a Community report, which provides a list of communities, separated by their HSDA and grouped by LHA, thereby allowing direct access to the community profile reports.

For each community, there is a provision to access a comprehensive community health printable report, which contains a collection of tables, charts, and other relevant information including: Historical Population Information (Figure 13), Health Indicators, Population Forecasts (with a focus on seniors) (Figure 14), Births, Immunization Information, Vaccine Preventable Diseases, Chronic Diseases, Senior Resident Profile (Figure 15), and Facility Activity and Available Services.

Figure 12. Community profile.

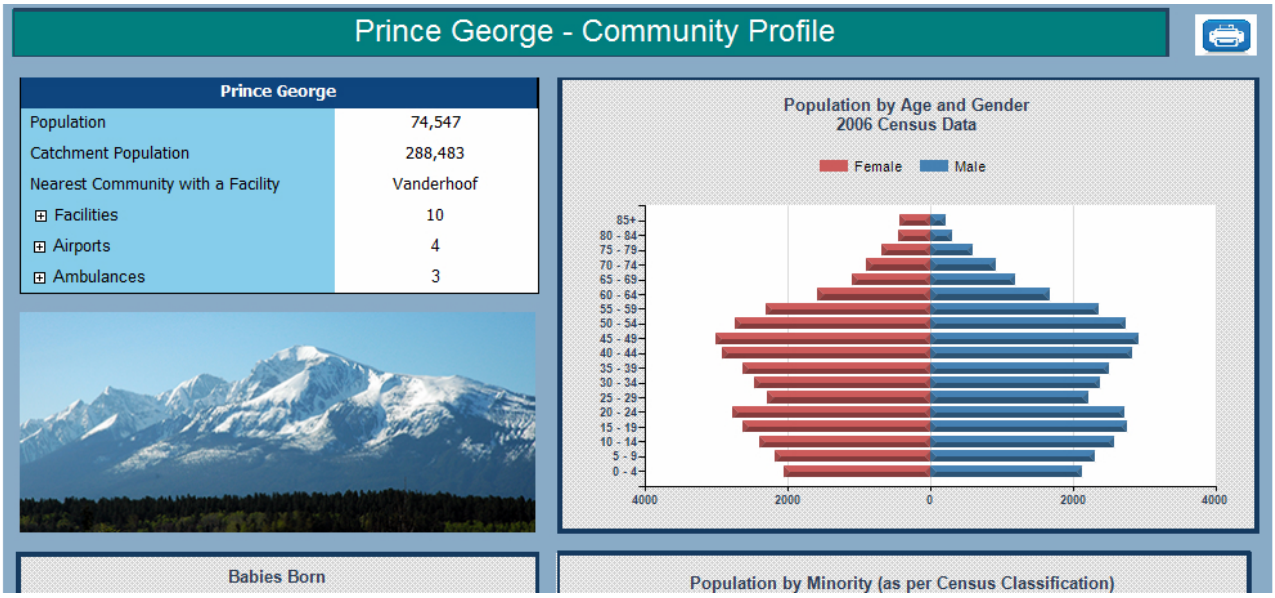
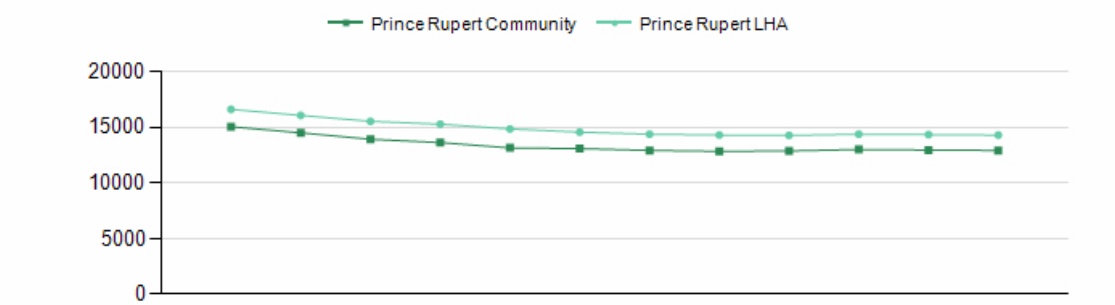


Figure 13. Population trend chart.



	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
C	15032	14485	13912	13610	13132	13072	12906	12831	12860	12989	12943	12913
LHA	16594	16052	15516	15266	14834	14547	14366	14287	14258	14366	14328	14284

Figure 14. Population projection.

Prince Rupert LHA: Population Projection						Population Change 2015-2030	
Broad Age Groups	2010	2015	2020	2025	2030	number	%
<20	3,833	3,587	3,518	3,446	3,341	-246	-6.9%
20-44	4,597	4,659	4,660	4,760	4,914	255	5.5%
45-64	4,300	4,338	4,284	3,918	3,577	-761	-17.5%
65+	1,629	1,988	2,554	3,220	3,698	1,710	86.0%
Total	14,359	14,572	15,016	15,344	15,530	958	6.6%
Focus on Seniors	2010	2015	2020	2025	2030	number	%
65+	1,629	1,988	2,554	3,220	3,698	1,710	86.0%
75+	645	711	890	1,171	1,508	797	112.1%
85+	175	214	260	294	363	149	69.6%
90+	76	88	125	139	153	65	73.9%

Figure 15. Senior residents profile.

Male	36 %	Married	23 %	Widowed	46 %
Aboriginal	23 %	Aged 75+	67 %	Female	64 %
Frequently Noted Health Conditions					
Diabetes	26 %	Hypertension	61 %	Osteoporosis/Cataract	22 %
Any Psychiatric Diagnosis	24 %	Arthritis	51 %	Chronic Arterial	24 %
Clients with Multiple Health Conditions					
3 - 5 Conditions	35 %	<=3 Conditions	25 %	>= 5 Conditions	40 %
Clients with independence difficulty in 1-3 daily activities (IADL Difficulty Scale)					
Great difficulty	17 %	No difficulty	33 %	Some difficulty	50 %
Clients with Cognitive Impairment (Cognitive Performace Score)					
Borderline /mild	42 %	Mod - Very Severe	6 %	No impairment	48 %

Other Reports

The reports illustrated in this paper are a small representative sample, due to space limitations. There are several other main and drill-down reports that provide various perspectives of the services availability. For instance, a community health report contains transfers/referrals information from/to the selected community in addition to charts and graphs that appear elsewhere in the application. These reports are printable and generally made available to communities. Similarly, many charts open a popup window instead of loading another report. These popups windows consist of descriptive charts or tables, definitions, and contain information about the source of data together with names of data analysts responsible for the information.

Discussion

Principal Findings

We have demonstrated how BI techniques and tools can be used in non-traditional areas of the health care environment to make informed decisions with reference to resource allocation and enhancement of the quality of patient care. The multidimensional cube allows analysis of data in several dimensions and reports are generated within seconds. The data can be kept up to date year round while preserving integrity during interim reporting. Originally, the data was updated annually due to the complexity of data collection and compilation. The versatility of reports is enhanced through parameterization, which allows values to be passed between sub-reports. The interaction of Web forms with the underlying database and cube allows for transparent data upload and integrity checks. The interactive reports provide users with valuable information such as proximity to location of available services, facilities with specific needs, comparative analysis, and tools for resource reallocation, if necessary. For privileged information, access controls have been implemented. The rural setting made this work more challenging because of the sparse geography and distance/travel times between facilities. Further, not all services are available in all communities, which requires identification of next best facility for repatriation of patients.

Conclusions

The overall impact of the work presented in this paper spans a number of areas such as better allocation of available funds and better outcomes by making informed decisions regarding medical and personnel resource utilization. Though these benefits have not been quantified, it has already been observed that analysts' time is now redirected to more effective surveillance activities and performance monitoring instead of collating data to manually generate reports.

It should also be noted that though the developed dashboard is not intended for real-time data, periodic surveillance reports can be generated on demand. Further, while the concept is applicable to all health authorities, it will be a challenge to have all jurisdictions collaborate and agree on a common architecture and/or report structures. Currently, the health planners and service providers internal to Northern Health are using the dashboard and planning is underway to have it accessible to the general public. The solution is modular and new datasets such as for smoking rates, teen pregnancies, HIV rates, immunization coverage, and vital statistical summaries can be easily integrated into the existing dashboard. The model can also be extended to other programs such as Home and Community Care, and Mental Health and Addictions. The next phase of this research is to determine how to incorporate services provided by non-NH providers such as Aboriginal Health Services.

Acknowledgments

This work was funded by a collaborative research grant from Northern Health, British Columbia, Canada. Among others, Kari Harder, James Haggerstone, Keely Maxwell, and Matthew Amsel have been very instrumental in compilation, loading, and verification of underlying data.

Conflicts of Interest

None declared.

References

1. Microsoft Corporation: Business Intelligence. URL: http://www.microsoft.com/en-us/server-cloud/solutions/business-intelligence/default.aspx [accessed 2014-05-30] [WebCite Cache ID 6Py0q6P30]
2. Microsoft Corporation: The Offical Microsoft ASP.NET Site. URL: http://www.asp.net/ [accessed 2014-05-30] [WebCite Cache ID 6Py1WHokZ]
3. Larson B. Delivering Business Intelligence with Microsoft SQL Server 2008. United States: McGraw-Hill; 2009.
4. Cantrill SV. Computers in patient care. Commun ACM 2010 Sep 01;53(9):42. [doi: 10.1145/1810891.1810907]
5. Pine M, Sonneborn M, Schindler J, Stanek M, Maeda JL, Hanlon C. Harnessing the power of enhanced data for healthcare quality improvement: lessons from a Minnesota Hospital Association Pilot Project. J Healthc Manag 2012;57(6):406-18; discussion 419. [Medline: 23297607]
6. Haque W, Edwards J. Ambulatory Care Sensitive Conditions: A business intelligence perspective. 2012 Presented at: Advances in Health Informatics Conference (AHIC); April 25, 2012; Toronto, Canada p. 31-39.
7. Olszak C, Batko K. The use of business intelligence systems in healthcare organizations in Poland. : IEEE; 2012 Presented at: Federated Conference on Computer Science and Information Systems (FedCSIS); September 9, 2012; Wroclaw, Poland p. 969-976.
8. Starfield B, Shi L, Macinko J. Contribution of primary care to health systems and health. Milbank Q 2005;83(3):457-502 [FREE Full text] [doi: 10.1111/j.1468-0009.2005.00409.x] [Medline: 16202000]
9. Lasser KE, Himmelstein DU, Woolhandler S. Access to care, health status, and health disparities in the United States and Canada: results of a cross-national population-based survey. Am J Public Health 2006 Jul;96(7):1300-1307. [doi: 10.2105/AJPH.2004.059402] [Medline: 16735628]
10. Sanmartin C, Berthelot JM, Ng E, Murphy K, Blackwell DL, Gentleman JF, et al. Comparing health and health care use in Canada and the United States. Health Aff (Millwood) 2006;25(4):1133-1142 [FREE Full text] [doi: 10.1377/hlthaff.25.4.1133] [Medline: 16835196]
11. Lurie N, Dubowitz T. Health disparities and access to health. JAMA 2007 Mar 14;297(10):1118-1121. [doi: 10.1001/jama.297.10.1118] [Medline: 17356034]
12. Statistics Canada. Population estimates and projections. URL: http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo52c-eng.htm [accessed 2014-07-15] [WebCite Cache ID 6R5xmLtCu]
13. BC Stats. Translations and data sets. URL: http://www.bcstats.gov.bc.ca/StatisticsBySubject/Geography/TranslationsDataSets.aspx [accessed 2014-05-30] [WebCite Cache ID 6Py1vOfW6]
14. QGIS Development Team. Quantum Geographic Information System. URL: http://www.qgis.org/en/site/ [accessed 2014-05-30] [WebCite Cache ID 6Py1odjM1]
15. Mäntyselkä P, Halonen P, Vehviläinen A, Takala J, Kumpusalo E. Access to and continuity of primary medical care of different providers as perceived by the Finnish population. Scand J Prim Health Care 2007 Mar;25(1):27-32 [FREE Full text] [doi: 10.1080/02813430601061106] [Medline: 17354156]

Abbreviations

BC: British Columbia
BI: business intelligence
CMG: case mix groups
ETL: extract-transform-load
GIS: geographic information system
HSDA: health service delivery area
KPI: key performance indicator
LHA: local health authority
NH: Northern Health
NHA: Northern Health Authority
OLAP: online analytical processing
SSIS: SQL Server Integration Services

Edited by C Pagliari; submitted 08.06.14; peer-reviewed by K Stroetmann, A Prins; comments to author 08.07.14; revised version received 15.07.14; accepted 18.07.14; published 06.08.14.

Please cite as:
Haque W, Urquhart B, Berg E, Dhanoa R
Using Business Intelligence to Analyze and Share Health System Infrastructure Data in a Rural Health Authority
JMIR Med Inform 2014;2(2):e16
URL: http://medinform.jmir.org/2014/2/e16/
doi:10.2196/medinform.3590
PMID:25599727

Original Paper

Design and Development of a Linked Open Data-Based Health Information Representation and Visualization System: Potentials and Preliminary Evaluation

Binyam Tilahun[1], MSc, MPH; Tomi Kauppinen[2], PhD; Carsten Keßler[3], PhD; Fleur Fritz[1], PhD

[1]Institute for Medical Informatics, University of Münster, Münster, Germany

[2]Cognitive Systems, University of Bremen, Bremen, Germany

[3]Center for Advanced Research on Spatial Information and Department of Geography, Hunter College, City University of New York, New York, NY, United States

Corresponding Author:
Binyam Tilahun, MSc, MPH
Institute for Medical Informatics
University of Münster
Albert-Schweitzer-Campus 1, Gebäude A11
Münster, D-48149
Germany
Phone: 49 (251) 83 55262
Fax: 49 (251) 83 52259
Email: Binyam.Tilahun@uni-muenster.de

Abstract

Background: Healthcare organizations around the world are challenged by pressures to reduce cost, improve coordination and outcome, and provide more with less. This requires effective planning and evidence-based practice by generating important information from available data. Thus, flexible and user-friendly ways to represent, query, and visualize health data becomes increasingly important. International organizations such as the World Health Organization (WHO) regularly publish vital data on priority health topics that can be utilized for public health policy and health service development. However, the data in most portals is displayed in either Excel or PDF formats, which makes information discovery and reuse difficult. Linked Open Data (LOD)—a new Semantic Web set of best practice of standards to publish and link heterogeneous data—can be applied to the representation and management of public level health data to alleviate such challenges. However, the technologies behind building LOD systems and their effectiveness for health data are yet to be assessed.

Objective: The objective of this study is to evaluate whether Linked Data technologies are potential options for health information representation, visualization, and retrieval systems development and to identify the available tools and methodologies to build Linked Data-based health information systems.

Methods: We used the Resource Description Framework (RDF) for data representation, Fuseki triple store for data storage, and Sgvizler for information visualization. Additionally, we integrated SPARQL query interface for interacting with the data. We primarily use the WHO health observatory dataset to test the system. All the data were represented using RDF and interlinked with other related datasets on the Web of Data using Silk—a link discovery framework for Web of Data. A preliminary usability assessment was conducted following the System Usability Scale (SUS) method.

Results: We developed an LOD-based health information representation, querying, and visualization system by using Linked Data tools. We imported more than 20,000 HIV-related data elements on mortality, prevalence, incidence, and related variables, which are freely available from the WHO global health observatory database. Additionally, we automatically linked 5312 data elements from DBpedia, Bio2RDF, and LinkedCT using the Silk framework. The system users can retrieve and visualize health information according to their interests. For users who are not familiar with SPARQL queries, we integrated a Linked Data search engine interface to search and browse the data. We used the system to represent and store the data, facilitating flexible queries and different kinds of visualizations. The preliminary user evaluation score by public health data managers and users was 82 on the SUS usability measurement scale. The need to write queries in the interface was the main reported difficulty of LOD-based systems to the end user.

Conclusions: The system introduced in this article shows that current LOD technologies are a promising alternative to represent heterogeneous health data in a flexible and reusable manner so that they can serve intelligent queries, and ultimately support decision-making. However, the development of advanced text-based search engines is necessary to increase its usability especially for nontechnical users. Further research with large datasets is recommended in the future to unfold the potential of Linked Data and Semantic Web for future health information systems development.

(JMIR Med Inform 2014;2(2):e31) doi:10.2196/medinform.3531

KEYWORDS

Linked Open Data; Semantic Web; ontology; health information systems; HIV; WHO; public health; public health informatics; visualization

Introduction

Information is a foundation for effective decision-making. This information need is even more critical in public health organizations to support areas such as epidemiologic surveillance, health outcome assessment, program evaluation and performance measurement, public health planning, and policy analysis [1]. In order to satisfy this, we need better and more flexible health data representation, analysis, querying, and visualization methods. The amount of available online health data both in structured and unstructured formats is constantly increasing. The World Health Organization (WHO), for example, has established a data repository providing access to over 50 datasets on priority health topics including mortality and prevalence of human immunodeficiency virus infection/acquired immunodeficiency syndrome (HIV/AIDS) in different WHO regions [2]. Moreover, the United Nations [3] and the Centers for Disease Control and Prevention (CDC) [4] have online data repositories on the different indicators for different countries.

While these are important initiatives to publish health data online, there has been relatively little attention paid to data representation methods in most health data portals so far [5]. Current data representation and distribution methods with only tabular formats, such as comma-separated values (CSV), PDF, and Excel—and little metadata—makes health information integration, comparison, and reuse very difficult. Additionally, even though different indicators have relationships to each other, the datasets are not linked in most portals. Vocabularies and data formats are inconsistent, which makes finding, assembling, and normalizing these datasets time consuming and prone to errors [6].

Exploiting the different kinds of public health information about a given topic is a challenging task because data is spread across different platforms in heterogeneous formats. Better data management methods and tools are required to move from a Web of documents, only understandable by human users, to a Web of Data in which information is expressed in a format that can be read and used by machines. This would enable us to find, share, and integrate information more easily [7].

Linked Data, as explained by Tim Berners-Lee [7], is a method to publish structured data by using standard Web technologies to connect related data and make them accessible on the Web. The Linked Data publishing pattern uses HTTP uniform resource identifiers (URIs) for identifying data items, the Resource Description Framework (RDF) for describing data, and links to describe the relationships. Other standards used in Linked Open Data (LOD) applications include Resource Description Framework Schema (RDFS) for describing RDF vocabularies, and SPARQL Protocol and RDF Query Language (SPARQL) for querying RDF graphs [8].

The primary goal of the Linked Data initiative is to make the World Wide Web (WWW) not only useful for interlinking documents, but also for sharing and interlinking data [9]. The movement is driven by the hypothesis that these technologies could revolutionize global data sharing, integration, and analysis, just like the classic Web-revolutionized information sharing and communication over the last two decades. However, to our knowledge there are not many studies on the potential of LOD for public health data management.

Motivated by the universal hypothesis of Linked Data to revolutionize data sharing, integration, and analysis, the main objectives of this work are (1) to test the potential of LOD for health data representation and visualization, (2) to identify the available technologies and tools for Linked Data-based health information system development, and (3) to evaluate the usability level of LOD-based systems by end users.

In this paper, we present the development of the system from data modeling to visualization and potential LOD tools available for development. Identifying the tools and testing the potential of LOD will be helpful as an input to the health informatics and Semantic Web community in the research effort to find ways to represent data in a flexible manner.

Methods

Overview

Our methodology was “Integration-oriented development and evaluation” in the sense that we used the available LOD tools to develop the system and then we reflected on the development process, the potentials, and finally on usability for end users. We gave special emphasis to the data management process as efficient data management and conversion is the backbone of the LOD-based system development [10]. We used the RDF for data representation, Fuseki triple store for data storage, and Sgvizler for information visualization. Additionally, we integrated a SPARQL query interface for interacting with the data. We primarily used the WHO health observatory dataset to test the system. All the data were represented using RDF and interlinked with other related datasets on the Web of Data using Silk [11], a link discovery framework for Web of Data. A

preliminary usability assessment was conducted following the System Usability Scale (SUS) method. The final revised SUS questionnaire used for the evaluation is shown in Multimedia Appendix 1. The details, with more focus on the data management process, are explained throughout this paper.

Data Sources

The dataset for this work was retrieved from the WHO global health observatory data repository [2]. The data used covered the years from 1990 to 2010. Missing data for some years were complemented with data from other similar official sources, such as the United Nations program for HIV/AIDS (UNAIDS) [3] and country-specific official sources like the national AIDS resource centers of each African country. From those databases, HIV statistical data, as well as additional location and total population information, were extracted for sub-Saharan African countries. Most of the data were in Microsoft Excel and CSV formats. All the data were converted and prepared in Excel using the Excel2RDF [12] converter. For the enrichment, DBpedia, Bio2RDF and LinkedCT were used as sources. For data license, all our published Linked Data adheres to the original data publisher's license and terms of use.

Data Modeling and Conversion

Shared vocabularies are a key to enable interoperability in healthcare systems by providing an agreed-upon terminology that can be looked up through URIs that cannot be referenced [13]. We have identified potential health, statistical, spatial, and time vocabularies and ontologies to share the data in a reusable way and then mapped them to the external ontologies using predicates (see Table 1). We used the common RDF [14], RDFS [15], Web Ontology Language (OWL) [16], friend of a friend (FOAF) [17], and Data Cube [18] vocabularies for data annotation. Those are standard vocabularies to represent data in LOD by expressing relationships between the data. We use the Data Cube vocabulary for all the statistical data to represent, not only the numbers, but also advanced metadata with space and time dimensions of the observation. Some of the standard predicates were replaced with more generic elements from the Data Cube vocabulary (eg, qb:prevalence instead of qb:observation) to make them more understandable to health professionals and healthcare managers. We assume that using some of the terms that are already known by health professionals will make the system more usable and easily adaptable. After identifying the ontologies and vocabularies, the original data was converted in a semi-automated way to avoid information loss. Conversion using Excel2RDF is done by selecting the range of data values and headers from the spreadsheet that are to be converted. Then, the headers are fed into the mapping wizard, which assists the mappings of row/column concepts to RDF vocabularies. Excel data triplication using Excel2RDF is discussed by Pesce et al [19].

Table 1. Different domain ontological vocabularies and predicates reused for modeling data in the conversion process.

Domain	Ontological vocabulary	Predicate
Health		
	prefix MeSH	Interlinked with <owl:sameAs>
	prefix Diseasome	Interlinked with <owl:sameAs>
	prefix dbpedia	Interlinkedwith<owl:sameAs>
Spatial		
	Prefix geo	geo:lat, geo:long
	prefix dcterm	dcterm:country
Statistical		
	prefix qb	qb:prevalence qb:slice qb:item
	prefix scovo	qb:prevalence qb:slice qb:item
Time series	prefix time	Time: year (from 1990-2010)

Data Storage

The main difference between existing health information system development and Linked Data-based systems is the way data is represented and stored. Current systems mostly use tabular formats (eg, Excel, CSV) or relational database systems such as Oracle. Linked Data-based systems, however, usually build on triple stores as their main data storage. This triple-based representation enables integration of data available from various sources without the need for physical storage of the RDF triple that corresponds to the relational data [20]. These systems provide data management and data access via application programming interfaces (APIs) and query languages to RDF data. For this work, we used the Fuseki triple store [21]. It provides representational state transfer (REST)-style SPARQL HTTP Update, SPARQL Query, and SPARQL Update using the SPARQL protocol over HTTP [22].

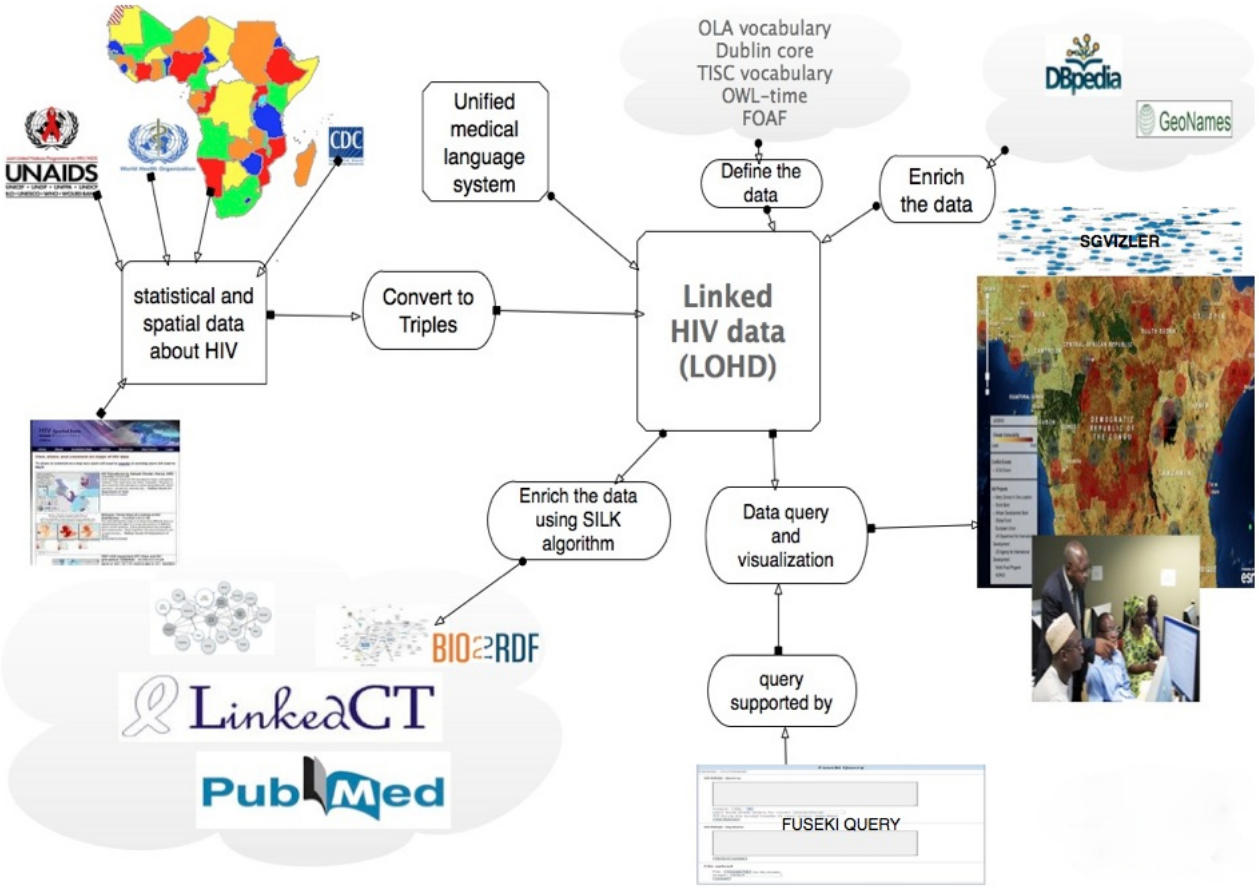
Data Enrichment

The primary intention of representing health data using the LOD approach is to be able to discover and link health data from different sources and use them in new applications. Interlinking data from our RDF datasets to other datasets, which are already in the LOD cloud, was challenging. It requires identification of similar link types in our datasets, and then finding suitable matching links in external datasets. Zevari et al point out similar challenges in link discovery in health datasets [23].

In our data enrichment, we used both manual and automatic methods. We manually enriched the dataset with links to some sources such as DBpedia, while large numbers of links to sources such as Bio2RDF were generated automatically. The enrichment is based on owl:sameAs relations, which interconnect different identifiers for the same real-world item across different datasets (eg, DBpedia:Ethiopia owl:sameAs geonames:7733022). Such a sameAs-link references different identifiers for the same real-world entity—Ethiopia, in our example—from different sources [10]. We enriched the data with links to data sources generated by related initiatives such as Bio2RDF [24], LinkedCT [25], Pubmed [26], and other geospatial and health-related initiatives using standard RDF and Unified Medical Language System (UMLS) vocabularies.

We used the Silk Link Discovery Framework [27] for automatic link discovery and to provide the built-in Fuseki query interface to access the data. To access the target data, we first configured access parameters to the target dataset endpoints using the <DataSource> directive. The only mandatory data source parameter is the endpoint URI. By specifying the source and destination endpoints on target datasets, we interlinked the data. In total, we retrieved 5312 data elements to be added to the system. Additionally, we implemented a visualization interface over the triple store using Sgvizler [28], a JavaScript library which renders the results of SPARQL queries as charts or HTML elements [29]. Figure 1 gives an overview of the overall methodology.

Figure 1. The overall workflow diagram for the methodology from the data conversion, data interlinking, and data query to visualization.



Results

Overview

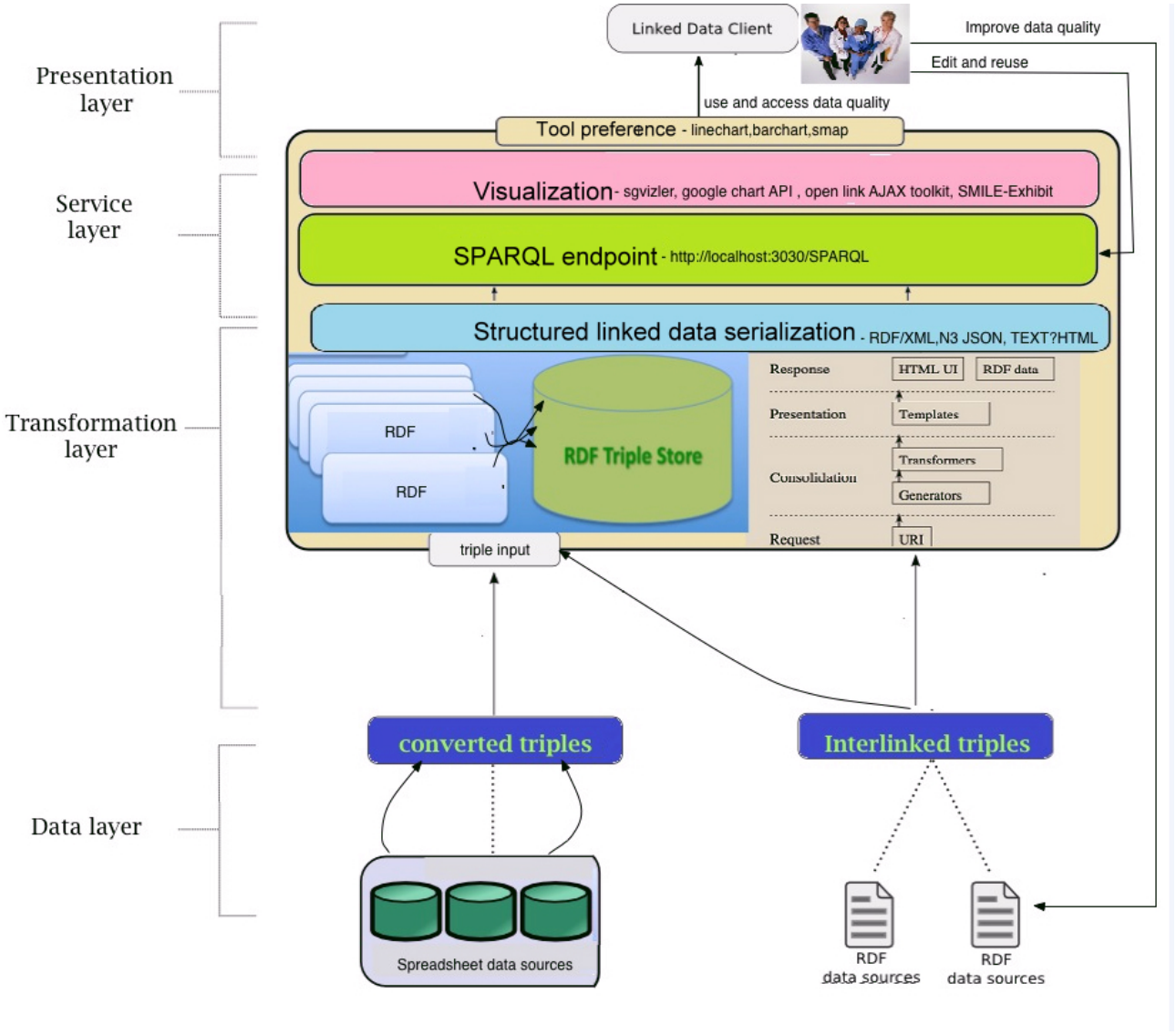
We developed a Linked Open Health Data (LOHD) system that integrates spatial and statistical health data from various sources. In the system, users can query HIV-related information about African countries and the system will support them in querying and visualizing the data in both space and time.

LOHD System Architecture

For the system development, we preferred a multilayer architecture, which provides flexibility and reusability. For example, data management, query processing, and visualization are logically separate processes. The advantages of a multilayer architecture have been discussed in the literature in detail [27-29]. By breaking up the system into a hierarchy, different layers can be developed sequentially and modified asynchronously without affecting the entire system architecture [10]. The architecture of our system is composed of 4 main layers (see Figure 2): (1) the data layer, (2) the transformation layer, (3) the service layer, and (4) the presentation layer. The data layer stores the converted and interlinked data. The transformation layer is the processing layer where every SPARQL query is processed using crawling pattern to localize data from the Web of Linked Data. The service layer controls the data access and bridges the client to the server via service protocols. The presentation layer allows the users to interact with the services using either retrieval or visualization tools.

All the system architecture layers and the underlying LOD application tools are shown in Figure 2.

Figure 2. LOHD System architecture with all the four layers from the data representation layer to presentation layer and the corresponding LOD tools.



Visualization

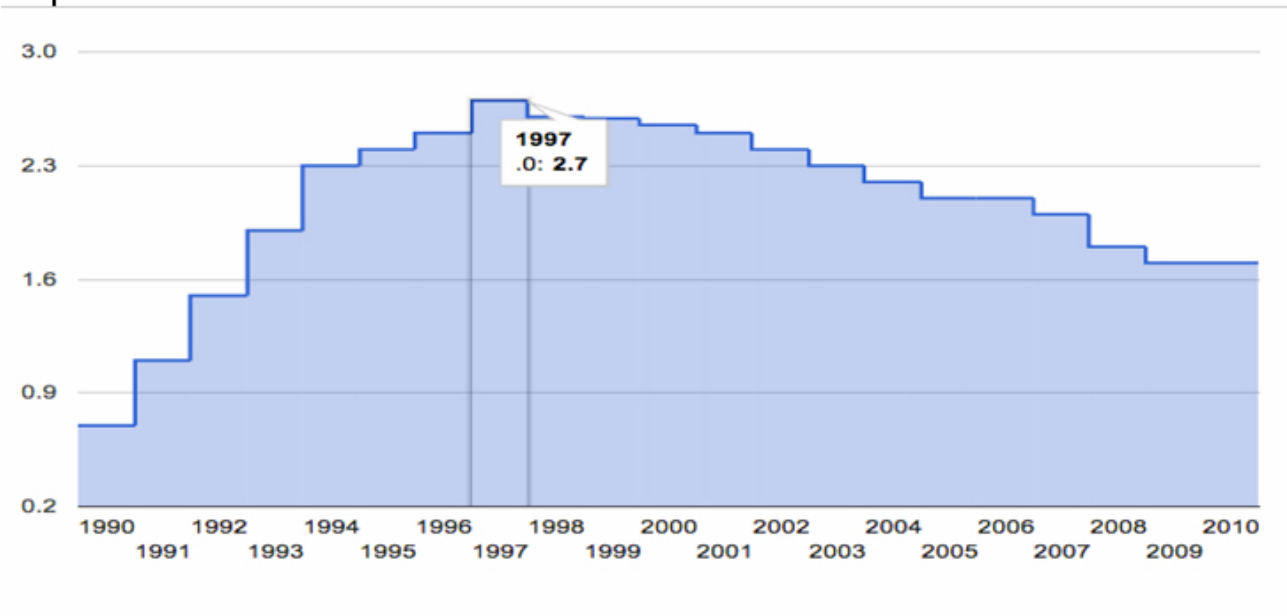
Coherent LOD visualizations enable nontechnical users to use the Web of Data [30] and increase the usability and accessibility of Linked Data-based systems [31]. In most Linked Data-based systems, the user is expected to write SPARQL queries, which is challenging for nontechnical users. To overcome those challenges, we integrate a live visualization interface using Sgvizler. Once the query is selected, the users have the option to choose the visualization method for the data output. All the visualization methods available on Sgvizler are supported by our system. In the following sample queries, we show some of the visualizations based on spatial or temporal queries.

Time Series Visualization of Linked Data

Time series visualizations help to display patterns and trends that are not readily apparent in the numbers themselves. In traditional databases, time series visualization is mostly done by external applications which are cumbersome and time consuming. But in Linked Data-based systems, you can write your query and choose the visualization type from the drop-down menu. Figure 3 shows the trend of HIV prevalence in Ethiopia, as an example, and the system automatically shows the live visualization of the trend for the requested year.

Figure 3. Time series visualization of HIV prevalence in Ethiopia from the years 1990-2010. To visualize other countries, substitute the country name in the query.

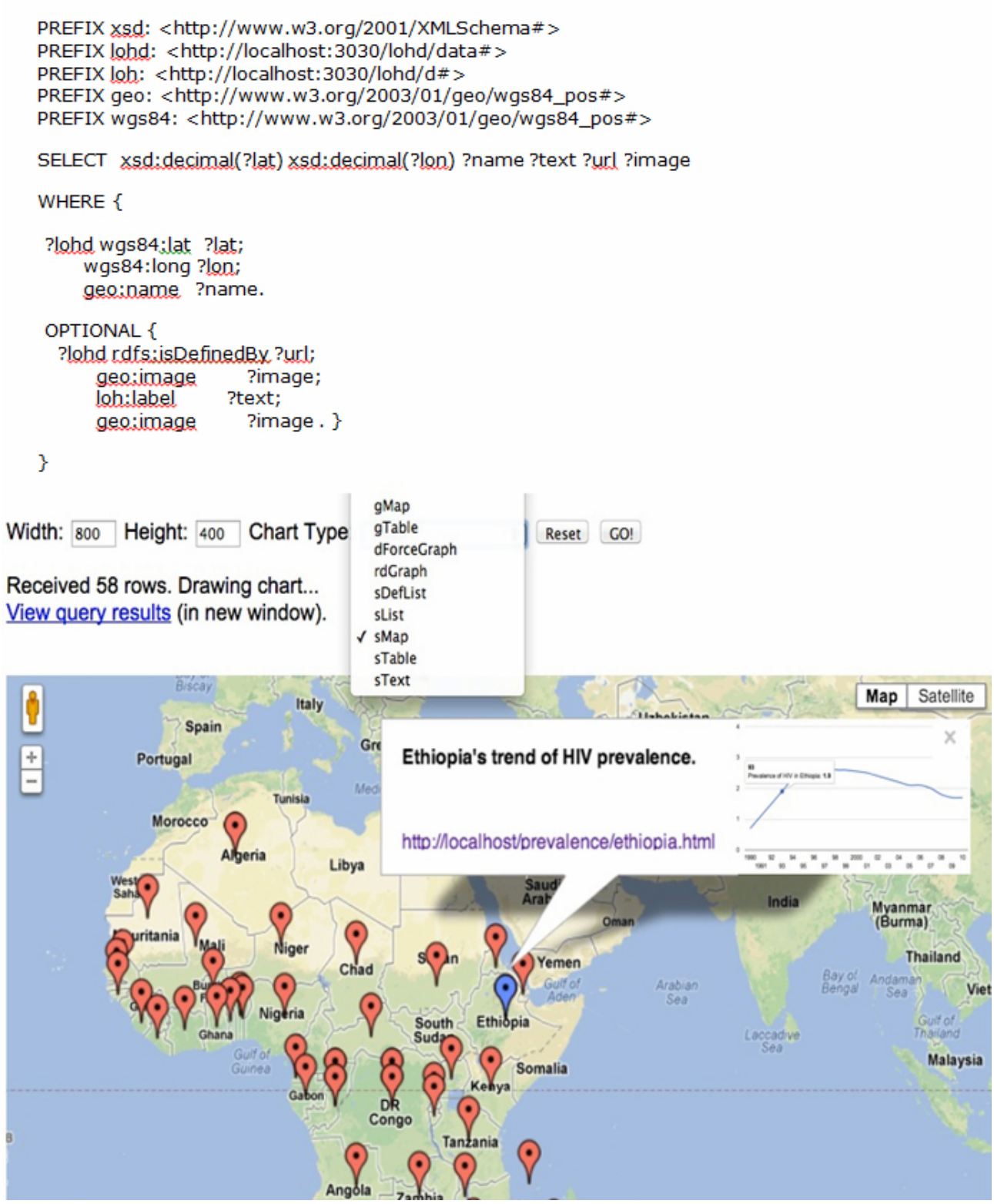
```
PREFIX lohd: <http://localhost:3030/lohd/data#>
PREFIX loh: <http://localhost:3030/lohd/d#>
PREFIX qb: http://purl.org/linked-data/cube#

SELECT ?year xsd:decimal(?prevalence)
where
{
?lohd lohd:year ?year;
 loh:countryname "Ethiopia";
 qb:prevalence ?prevalence.
}
order by ?year
```



Geographical Visualization of Linked Data

Location is becoming a basic attribute for health data [32]. Location-based visualizations are mostly difficult using traditional databases unless they are exported to geographic information system (GIS) software for further analysis. In LOD-based systems, location-based visualizations are facilitated by the ability to write queries and choose the visualization method. Figure 4 shows an example where the visualization shows the prevalence of HIV based on each country's location on the African map. When someone clicks on the icon of the country, it will show the basic information about the country and the trend of HIV for the specified time period in the query.

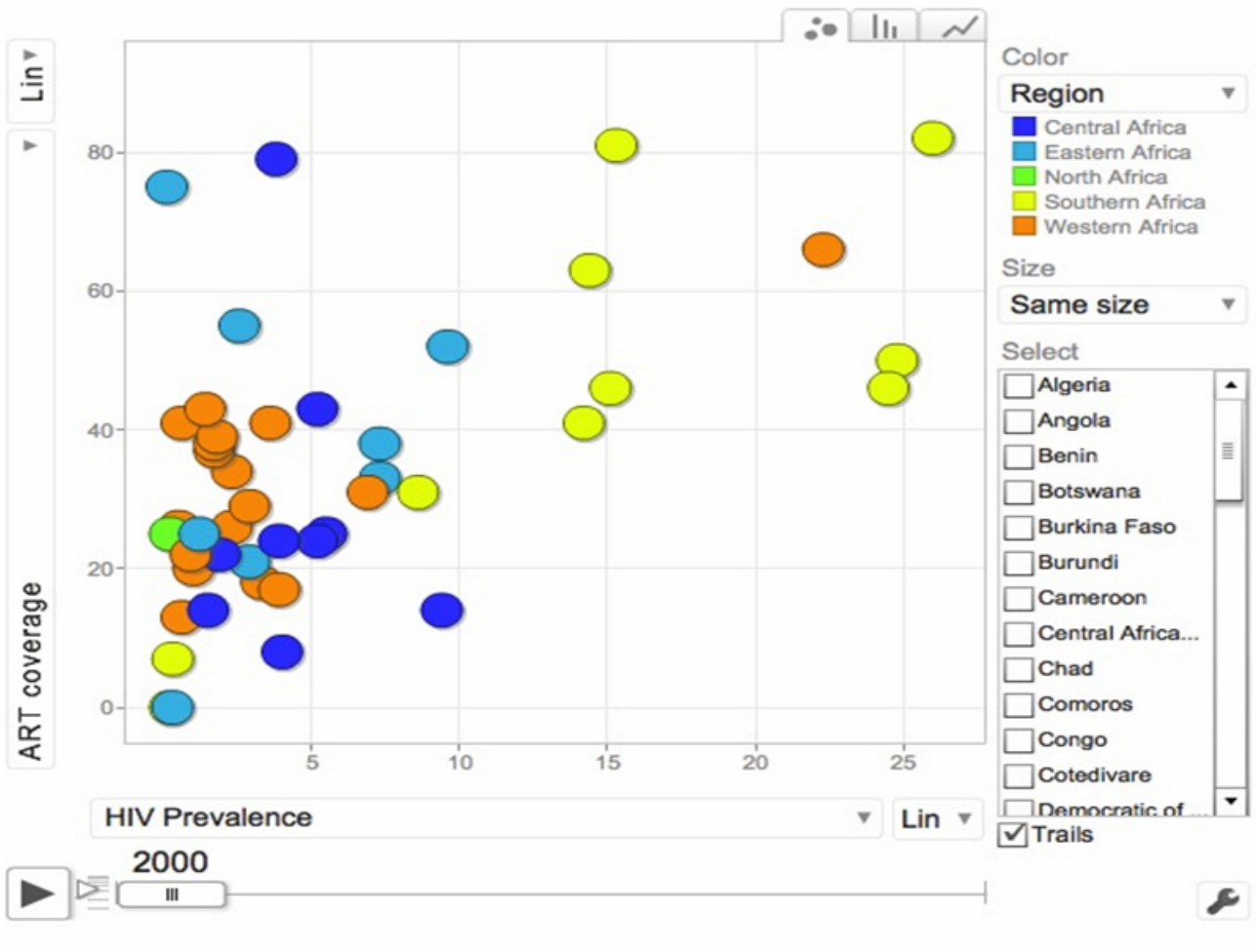
Figure 4. Location-based visualization of HIV prevalence in sub-Saharan Africa. The health-related data and the time series graph are displayed by clicking on the map of the country.



Indicator-Based Visualization of Linked Data

Indicators are the basic components of any health data. Most international disease prevalence comparisons and local-level reporting are done using indicators in a specific period of time. LOD-based systems support queries with different indicators—such as HIV prevalence rate by country or region, antiretroviral therapy (ART) coverage rate, population or gross domestic product (GDP)—and make a correlation analysis between those variables over time. In Figure 5, we show a 3-dimensional correlation analysis with time series animation.

Figure 5. Indicator-based correlation visualization over LOHD system of HIV prevalence and ART coverage versus time.



Evaluation

The system was evaluated in a small-scale user study to get feedback from healthcare data managers and users regarding the usability and learnability of the system. A total of 19 participants were selected for this evaluation, both with a technical and nontechnical background. The participants had no relationship with the investigator and the selection was done purposefully to ensure we recruited participants who currently work on health data management, and to get a proportional mix of different professions. Of the 19 selected participants, 17 of them responded to the questionnaire (89%). The technical participants (9/17, 53%) were data managers with IT backgrounds, health information system developers, and system administrators in different healthcare organizations in Africa. The nontechnical participants (8/17, 47%) were public data users, such as demographic data managers, doctors, and public health professionals. The evaluation was done based on the SUS with some wording amended, tailored for our participants (see Multimedia Appendix 1). In the evaluation, we were interested in the feedback from the participants on the query-based data access. The Linked Data search engine was not provided to participants, given its early stage of development for complex query request. The SUS is mostly regarded as a quick and easy way to conduct a usability assessment [33]. Even though the tool is self-described as “quick and dirty”, it has been evaluated in many studies (more than 600 articles) as valid and reliable [34]. Based on the SUS scoring criteria, the final calculated score was 82, which is well above the average SUS score of 68. Table 2 summarizes the evaluation responses for each criteria of the system usability.

Additionally, 2 open-ended questions were asked to the users to better understand their views and their specific requirements for using the system. The frequent answers for those questions can be explained by dividing them into 2 groups. The participants with technical backgrounds were relatively happy and 8 out of 9 (89%) of them mentioned that such systems would be useful in the future. The nontechnical participants (8/8), however, mentioned that the system was not easy to use. This is understandable seeing that the current Linked Data tools demand writing queries. Publishing the data in machine-understandable form and making live visualization without having to use external applications were the most frequently mentioned benefits of the system by the participants of the evaluation (15/17, 88%). The need to write queries in the optional interface and identifying the appropriate visualization tool were reported as being the difficult aspects of such systems by 16 of the 17 (94%) participants.

Table 2. . SUS evaluation criteria and participant response (n=17).

SUS evaluation criteria	Strongly agree, n (%)	Agree, n (%)	Neutral, n (%)	Disagree, n (%)	Strongly disagree, n (%)
I think that I would like to access my data this way.	10 (59)	3 (18)	-	4 (24)	-
I found the system unnecessarily complex.	2 (12)	5 (29)	-	10 (59)	-
I thought the system was easy to use.	4 (24)	5 (29)	-	7 (41)	1 (6)
I think that I would need the support of a technical person to be able to use this system.	6 (35)	1 (6)	-	7 (41)	3 (18)
I found the various functions in this system well integrated.	12 (71)	2 (12)	-	3 (18)	-
I thought there was too much inconsistency in this system.	3 (18)	2 (12)	-	10 (59)	1 (6)
I would imagine that most people would learn to use this system very quickly.	6 (35)	1 (6)	2 (12)	6 (35)	3 (18)
I found the system very cumbersome to use.	4 (24)	2 (12)	-	11 (65)	1 (6)
I felt very confident using the system.	12 (71)	-	-	5 (29)	-
I needed to learn a lot of things before I could get going with this system.	7 (41)	1 (6)	-	9 (53)	-

Discussion

Principal Findings

We developed a Linked Data-based health information representation, querying, and visualization system. We used the system to represent and store the data, facilitating flexible queries and different kinds of visualizations. There are other ongoing efforts to convert healthcare- and life science-related datasets to a Linked Data cloud such as Linked Open Drug Data (LODD), LinkedCT, Open Biomedical Ontologies (OBO), and the World Wide Web Consortium's (W3C) Health Care and Life Sciences working groups [31,32]. Thanks to such initiatives and recently developed Semantic Web tools, converting data to RDF has become straightforward. However, just converting the data to RDF and publishing it online is not enough [35,36]. The main difficulty is to integrate the data representation methods to application-level tools and make them usable for health information consumers in a shared, semantically meaningful, easily discoverable, and reusable manner.

In our system, we represented the health data with its important dimensions—magnitude, time, and space—in the form of RDF and we used both manual and automatic interconnection methods to enrich the data. We integrated visualization and retrieval methods for the data to make data visualization and retrieval possible with already available tools. There was a similar initiative by Zappa et al to integrate mutation data in the LOD cloud [35]. The methodology we follow for development is similar except that they use another tool for the data conversion. What makes our work different is that in addition to converting the data and making it available in RDF, we focus on integrating additional query and visualization interface tools to make the system more usable, especially for nontechnical users.

Our system development method was integration oriented in the sense that it reflects the way to convert the different dimensions of the data to Linked Data and integrate them with already developed tools, enabling the system to support information access. In selecting our tools, we found out that RDF is currently a robust data model to represent data with metadata [14,37] that gives the opportunity of integrating data and availing data for query. Our selection of Sgvizler for visualization was motivated by its current support of different types of visualization and its integration with HTML webpages by letting the user specify queries of interest [29]. One of the difficulties we noticed here is that for complex queries, Sgvizler is relatively slow. This may make it difficult to use for big data and complex query-based systems. Nonetheless, we believe that advanced-level, live correlation visualization of certain disease trends in space and time dimensions from different sources is one of the biggest promises of Linked Data-based systems in the future.

Measuring the degree of advancement that a Linked Data representation brings to public health data is difficult to quantify. Nonetheless, from the technology perspective, the data becomes search engine discoverable and machine understandable, which addresses the main issues of the current health data silos problem [38]. While Linked Data and Semantic Web technologies are not as mature as other database technologies, they present a promising alternative in public health information portal development. A good example that can explain this is the data representation scheme in the World Bank database [39], which includes both a portal for downloading data as Excel or PDF files, as well as a Linked Data version for downloading the data as RDF with the ability to query their endpoints. The main advantage of having Linked Data as an additional option in the World Bank database can be seen in the results of search engine results. If you input "Prevalence of HIV in Egypt" and "GDP of Egypt" into search engines, we can clearly see the data representation limitation of health portals. Since the World Bank data is represented in a machine-understandable and search engine-discoverable way, you can see the graphs and additional descriptions, which are very useful for an end user searching for them.

The user evaluation of our system confirms the existing usability limitations of Linked Data mentioned by different authors [21,32,35,36,40]. Linked Data is currently mostly used by the Semantic Web community and other users with a strong technical background. To make the Linked Data-based systems more usable by end users, we need to develop enhanced tools that can avoid the need to write queries.

In our evaluation, 41% (7/17) of the participants (strongly agree and agree together) reported that they need the support of a technical person to use this system, which is high when compared to other system evaluations [4,33]. Yet this is an expected result given the current technical nature of data access in LOD when using queries. The promising result from the evaluation is that 70% (12/17) of the participants are confident in using and understanding the visualizations of the system. This indicates that the LOD-based representation of public health data offers a new perspective in the future of health data portal development.

Limitations

There are some limitations in this work. Primarily, the amount of data we used is small to generalize the robustness of the LOD tools. As already outlined in different studies [41-43] Semantic Web technologies work well with small datasets but might not be the best option with big datasets. Secondly, our user evaluation was based on a small set of participants and the SUS scale, which has its own limitations, making generalization of the usability assessment result difficult.

For future research we recommend integrating and testing an advanced-level search engine to ensure that LOD-based systems are more usable outside the Semantic Web community. Additionally, implementing and testing a similar system with a big dataset by describing the data more robustly with domain-specific, additional ontological vocabularies, interlinking with more ontologies, and including more visualization options for grouped data is recommended. Moreover, implementation of advanced-level correlation analysis visualization from different sources will make LOD technology more interesting and usable by healthcare professionals.

Conclusions

The system introduced in this article shows that LOD has a promising potential in the representation of complex health-related data. This is mainly due to its reusable and interoperable manner that can serve intelligent queries, and ultimately support decision-making. However, the development of advanced LOD search engines is necessary to increase its usability.

Acknowledgments

We thank the WHO for publishing the data online and for making it freely available. We also want to thank the participants of the evaluation and the anonymous reviewer who gave us important comments for the improvement of the paper.

Authors' Contributions

BT developed and implemented the system, guided the evaluation and wrote the manuscript. CK and TK contributed to the study design and in critically revising the manuscript. FF contributed to the manuscript design and to the critical evaluation of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Final revised SUS questionnaire for the evaluation.

[PDF File (Adobe PDF File), 102KB - medinform_v2i2e31_app1.pdf]

References

1. Studnicki J, Berndt DJ, Fisher JW. Using information systems for public health administration. In: Novick LF, Morrow CB, Mays GP, editors. Public Health Administration: Principles for Population-Based Management. 2nd edition. Sudbury, MA: Jones and Bartlett; 2008:353-380.
2. World Health Organization. 2011. Global Health Observatory URL: http://www.who.int/gho/database/en/ [accessed 2014-04-15] [WebCite Cache ID 6Or9H1Dzq]
3. UNAIDS. 2012. UNAIDS HIV database and visualization URL: http://www.unaids.org/en/dataanalysis/ [accessed 2014-04-15] [WebCite Cache ID 6OrA8yElt]
4. Centers for Disease Control and Prevention. 2012. HIV/AIDS database URL: http://www.cdc.gov/hiv/statistics/basics/ [accessed 2014-04-15] [WebCite Cache ID 6OrAKhX0c]
5. Gao S, Mioc D, Yi X, Anton F, Oldfield E, Coleman DJ. Towards Web-based representation and processing of health information. Int J Health Geogr 2009;8:3 [FREE Full text] [doi: 10.1186/1476-072X-8-3] [Medline: 19159445]

6. Battista ADL, Villanueva-Rosales N, Palenychka M, Dumontier M. SMART: A Web-based, ontology-driven, Semantic Web query answering application. In: Proceedings of the Semantic Web Challenge. Spain: CEUR Workshop Proceedings; 2007 Presented at: Semantic Web Challenge; November 13, 2007; Busan, South Korea URL: http://ceur-ws.org/Vol-295/paper17.pdf
7. Berners-Lee T. World Wide Web Consortium. 2009. Linked data URL: http://www.w3.org/DesignIssues/LinkedData.html [accessed 2014-08-04] [WebCite Cache ID 6RZsqFSSL]
8. Hitzler P, Janowicz K. Linked data, big data, and the 4th paradigm. Semant Web 2013;4(3):233-235 [FREE Full text]
9. Bizer C, Heath T, Berners-Lee T. Linked data - the story so far. Int J Semant Web Inf Syst 2009 Feb 12;5(3):1-22 [FREE Full text] [doi: 10.4018/jswis.2009081901]
10. Gür N, Sanchez LD, Kauppinen T. GI systems for public health with an ontology-based approach. In: Proceedings of the AGILE'2012 International Conference on Geographic Information Science. 2012 Presented at: AGILE'2012 International Conference on Geographic Information Science; April 24-27, 2012; Avignon, France p. 86-91 URL: http://www.agile-online.org/Conference_Paper/CDs/agile_2012/proceedings/papers/Paper_Guer_GI_Systems_for_Public_Health_with_an_Ontology_Based_Approach_2012.pdf
11. Isele R, Jentzsch A, Bizer C, Volz J, Petrovski P. University of Mannheim. Silk: A link discovery framework for the Web of Data URL: http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/ [accessed 2014-08-04] [WebCite Cache ID 6RZtgQVoQ]
12. Mohammad. GitHub. 2013 Jun 04. Excel2rdf plugin URL: https://github.com/waqarini/excel2rdf [accessed 2014-05-08] [WebCite Cache ID 6PPucOY96]
13. Tao C, Jiang G, Wei W, Solbrig HR, Chute CG. Towards Semantic Web-based representation and harmonization of standard meta-data models for clinical studies. AMIA Jt Summits Transl Sci Proc 2011;2011:59-63 [FREE Full text] [Medline: 22211181]
14. World Wide Web Consortium.: RDF Working Group; 2011. Resource description framework (RDF) URL: http://www.w3.org/RDF/ [accessed 2014-05-08] [WebCite Cache ID 6PPv2xoiu]
15. Schema.org. 2012. What is Schema.RDFS.org? URL: http://schema.rdfs.org/ [accessed 2014-05-08] [WebCite Cache ID 6PPv9YagB]
16. World Wide Web Consortium. 2013. OWL Web ontology language current status URL: http://www.w3.org/standards/techs/owl [accessed 2014-08-04] [WebCite Cache ID 6RZvBMyon]
17. xmlns.com. 2014 Jan 14. FOAF vocabulary specification 0.99 URL: http://xmlns.com/foaf/spec/ [accessed 2014-05-08] [WebCite Cache ID 6PPvTt1Th]
18. Tennison J. World Wide Web Consortium. 2014 Jan 16. RDF data cube vocabulary URL: https://github.com/waqarini/excel2rdf [accessed 2014-10-15] [WebCite Cache ID 6TMFyPB5Q]
19. Pesce ML, Breitman KK, Casanova MA. Surfacing scientific and financial data with the Xcel2RDF plug-in. In: 2nd Workshop on Developing Tools as Plug-Ins (TOPI).: IEEE; 2012 Presented at: IEEE Conference on Developing Tools as Plug-Ins (TOPI); June 3, 2012; Zurich, Switzerland p. 73-78. [doi: 10.1109/TOPI.2012.6229814]
20. Oracle Spatial and Graph RDF Semantic Graph Developer's Guide. 2013. 10 RDF views: relational data as RDF URL: http://docs.oracle.com/database/121/RDFRM/sem_relational_views.htm [accessed 2014-08-04] [WebCite Cache ID 6RZvysBeU]
21. Apache Jena. 2013. Fuseki: serving RDF data over HTTP URL: http://jena.apache.org/documentation/serving_data/ [accessed 2014-05-08] [WebCite Cache ID 6PPvdqBZH]
22. Hu Y, Janowicz K, McKenzie G, Sengupta K, Hitzler P. A Linked-Data-driven and semantically-enabled journal portal for scientometrics. Lecture Notes in Computer Science 2013;8219:114-129. [doi: 10.1007/978-3-642-41338-4_8]
23. Zaveri A, Lehmann J, Auer S, Hassan MM, Sherif MA, Martin M. Publishing and interlinking the Global Health Observatory dataset: towards increasing transparency in global health. Semant Web 2012;1.
24. Callahan A, Cruz-Toledo J, Dumontier M. Ontology-based querying with Bio2RDF's Linked Open Data. J Biomed Semantics 2013 Apr 15;4 Suppl 1:S1 [FREE Full text] [doi: 10.1186/2041-1480-4-S1-S1]
25. Hassanzadeh O, Kementsietsidis A, Lim L, Miller RJ, Wang M. LinkedCT: A linked data space for clinical trials. The Computing Research Repository (CoRR) 2009 Aug 04;abs/0908.0.
26. Yamamoto Y, Yamaguchi A, Yonezawa A. Building Linked Open Data towards integration of biomedical scientific literature with DBpedia. J Biomed Semantics 2013;4(1):8 [FREE Full text] [doi: 10.1186/2041-1480-4-8] [Medline: 23497538]
27. Jentzsch A, Isele R, Bizer C. Silk: Generating RDF links while publishing or consuming Linked Data. In: CEUR Workshop Proceedings. 2010 Presented at: International Semantic Web Conference; November 9, 2010; Shanghai, China.
28. Skjæveland MG. Sgvizler: A JavaScript wrapper for easy visualization of SPARQL result sets. 2012 Presented at: 9th Extended Semantic Web Conference; 2012; Heraklion, Crete, Greece.
29. Skjæveland MG. Data 2000. Sgvizler URL: https://code.google.com/p/sgvizler/ [accessed 2014-05-08] [WebCite Cache ID 6PPwJLJYF]
30. Brunetti JM, Auer S, García R. The Linked Data Visualization Model. In: CEUR Workshop Proceedings. 2012 Presented at: International Semantic Web Conference; November 11-15, 2012; Boston, USA.

31. Kopanitsa G, Hildebrand C, Stausberg J, Englmeier KH. Visualization of medical data based on EHR standards. Methods Inf Med 2013;52(1):43-50. [doi: 10.3414/ME12-01-0016] [Medline: 23223709]
32. Andes N, Davis JE. Linking public health data using geographic information system techniques: Alaskan community characteristics and infant mortality. Stat Med 1995;14(5-7):481-490. [Medline: 7792442]
33. Sauro J. MeasuringU. Denver, CO; 2011 Feb 02. Measuring usability with the System Usability Scale (SUS) URL: http://www.measuringusability.com/sus.php [accessed 2014-05-08] [WebCite Cache ID 6PPwyltyj]
34. Mazumdar S, Petrelli D, Ciravegna F. Exploring user and system requirements of Linked Data visualization through a visual dashboard approach. Semant Web 2011 Nov 20:1-18.
35. Zappa A, Splendiani A, Romano P. Towards linked open gene mutations data. BMC Bioinformatics 2012;13 Suppl 4:S7 [FREE Full text] [doi: 10.1186/1471-2105-13-S4-S7] [Medline: 22536974]
36. Eysenbach G. The Semantic Web and healthcare consumers: a new challenge and opportunity on the horizon? IJHTM 2003;5(3/4/5):194-212. [doi: 10.1504/IJHTM.2003.004165]
37. Reynolds D. World Wide Web Consortium. 2010. Data Cube implementations URL: http://www.w3.org/2011/gld/wiki/Data_Cube_Implementations [accessed 2014-08-04] [WebCite Cache ID 6RZyJX4sX]
38. Semple H, Qin H, Sasson C. Development of a Web GIS application for visualizing and analyzing community out of hospital cardiac arrest patterns. Online J Public Health Inform 2013;5(2):212 [FREE Full text] [doi: 10.5210/ojphi.v5i2.4587] [Medline: 23923097]
39. World Bank Linked Data. 2013. Observations in World Bank URL: http://worldbank.270a.info/view [accessed 2014-08-04] [WebCite Cache ID 6RZyWPNDF]
40. Samwald M, Jentzsch A, Bouton C, Kallesøe CS, Willighagen E, Hajagos J, et al. Linked open drug data for pharmaceutical research and development. J Cheminform 2011 May 16;3(1):19 [FREE Full text] [doi: 10.1186/1758-2946-3-19] [Medline: 21575203]
41. Bukhari AC, Baker CJO. The Canadian health census as Linked Open Data: towards policy making in public health. 2013 Presented at: 9th International Conference on Data Integration in the Life Sciences; July 11-12, 2013; Montreal, PQ URL: http://www2.unb.ca/csas/data/ws/dils2013/papers/DILS-SYS-EC-paper3.pdf
42. Pathak J, Kiefer RC, Chute CG. The linked clinical data project: applying Semantic Web technologies for clinical and translational research using electronic medical records. In: Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences. New York, NY, USA: ACM; 2012 Presented at: 4th International SWAT4LS Workshop; December 7-9, 2011; London, UK p. 94-95 URL: http://www.jbiomedsem.com/content/3/1/10 [doi: 10.1186/2041-1480-3-10]
43. Tao C, Jiang G, Oniki TA, Freimuth RR, Zhu Q, Sharma D, et al. A Semantic-Web oriented representation of the clinical element model for secondary use of electronic health records data. J Am Med Inform Assoc 2013 May 1;20(3):554-562 [FREE Full text] [doi: 10.1136/amiajnl-2012-001326] [Medline: 23268487]

Abbreviations

API: application programming interface
ART: antiretroviral therapy
CDC: Centers for Disease Control and Prevention
FOAF: friend of a friend
GDP: gross domestic product
GIS: geographic information system
LOD: Linked Open Data
OWL: Web Ontology Language
RDF: Resource Description Framework
RDFS: Resource Description Framework Schema
REST: representational state transfer
SPARQL: SPARQL Protocol and RDF Query Language
SUS: System Usability Scale
UMLS: Unified Medical Language System
URI: uniform resource identifier
W3C: World Wide Web Consortium
WHO: World Health Organization

Edited by I Buchan; submitted 10.05.14; peer-reviewed by N Peek; comments to author 08.07.14; revised version received 04.08.14; accepted 23.08.14; published 25.10.14.

Please cite as:
Tilahun B, Kauppinen T, Keßler C, Fritz F
Design and Development of a Linked Open Data-Based Health Information Representation and Visualization System: Potentials and Preliminary Evaluation
JMIR Med Inform 2014;2(2):e31
URL: http://medinform.jmir.org/2014/2/e31/
doi:10.2196/medinform.3531
PMID:25601195

Viewpoint

Towards Social Radiology as an Information Infrastructure: Reconciling the Local With the Global

Gustavo Henrique Matos Bezerra Motta[1], CSci, PhD

Centro de Informática, Departamento de Informática, Universidade Federal da Paraíba, João Pessoa, Brazil

Corresponding Author:
Gustavo Henrique Matos Bezerra Motta, CSci, PhD
Centro de Informática
Departamento de Informática
Universidade Federal da Paraíba
Jardim Universitário, s/n - Castelo Branco
João Pessoa, 58051-900
Brazil
Phone: 55 83 3216 7093
Fax: 55 83 3216 7093
Email: gustavo@ci.ufpb.br

Abstract

The current widespread use of medical images and imaging procedures in clinical practice and patient diagnosis has brought about an increase in the demand for sharing medical imaging studies among health professionals in an easy and effective manner. This article reveals the existence of a polarization between the local and global demands for radiology practice. While there are no major barriers for sharing such studies, when access is made from a (local) picture archive and communication system (PACS) within the domain of a healthcare organization, there are a number of impediments for sharing studies among health professionals on a global scale. Social radiology as an information infrastructure involves the notion of a shared infrastructure as a public good, affording a social space where people, organizations and technical components may spontaneously form associations in order to share clinical information linked to patient care and radiology practice. This article shows however, that such polarization establishes a tension between local and global demands, which hinders the emergence of social radiology as an information infrastructure. Based on an analysis of the social space for radiology practice, the present article has observed that this tension persists due to the inertia of a locally installed base in radiology departments, for which common teleradiology models are not truly capable of reorganizing as a global social space for radiology practice. Reconciling the local with the global signifies integrating PACS and teleradiology into an evolving, secure, heterogeneous, shared, open information infrastructure where the conceptual boundaries between (local) PACS and (global) teleradiology are transparent, signaling the emergence of social radiology as an information infrastructure.

(JMIR Med Inform 2014;2(2):e27) doi:10.2196/medinform.3648

KEYWORDS

information infrastructures; social radiology; teleradiology; picture archive and communication systems (PACS); sociotechnical systems

Introduction

Contemplating social radiology in terms of an information infrastructure [1-7] goes beyond discussion on technologies for archiving and transmitting medical images or tools for medical imaging. It involves the notion of a shared infrastructure as a public good [2], capable of supporting the formation of associations between people, organizations and technical components in order to support patient care and radiology practice in a networked world. The emergence and sustainability of an information infrastructure require a permanent endeavor in order to build and maintain a set of solutions distributed along the social or technical and local or global axes of the information infrastructure space [2]. When solutions polarize, emphasizing the local (technical) aspect rather than the global (social), or vice-versa, the information infrastructure does not emerge. In a recent work [5], the term “knowledge infrastructure” is used as new terminology for “information infrastructure”.

This article reveals the existence of a polarization between the local and global demands for radiology practice, thus jeopardizing the emergence of social radiology as an information infrastructure. The article begins by demonstrating that such

polarization establishes a tension between local and global radiology practices and that this tension denotes the lack of an information infrastructure. The following explores the concept of an information infrastructure for radiology practice as a social space that is interactive, evolving, and open. Next, it analyses how the struggle with the inertia of the installed base impedes the emergence of social radiology as an information infrastructure. Additionally, it also argues that current teleradiology models do not configure as an information infrastructure for social radiology. The article concludes by discussing the manner in which reconciliation between local and global is a way towards social radiology as an information infrastructure.

The Tension Between Local and Global Radiology Practices

Sharing medical imaging studies among health professionals in an easy and effective manner has long been an on-going pursuit in radiology [8]. This is even more evident nowadays, with the widespread use of medical images and imaging procedures in clinical practice and patient diagnosis. The demand for noninvasive diagnostic imaging tests continues to increase, where the growing trend among non-radiologist physicians is twice as fast as among radiologists [9]. As such, the timely access to medical imaging studies by radiologists and non-radiologists is imperative.

In general, radiologists have no major issues with reading images and creating primary diagnostic reports. Similarly, other health professionals have no concerns about reading such images and reports when access is within the domain of a health care organization. Image related data and working functions are accessed in a workflow supported by the picture archive and communication systems (PACS) and radiology information systems (RIS) of a radiology department [10,11]. When access is required from a remote location, health care organizations typically adopt a suitable teleradiology solution. For instance, they adopt virtual private networks (VPN) and cloud computing technologies to enable physicians to access PACS/RIS from a different location or to integrate geographically separated buildings within a health care organization [12-14]. Another solution commonly used is outsourcing image interpretation services. In this case, regional, national or international teleradiology companies only interpret or broker image interpretation for non-radiologists [15].

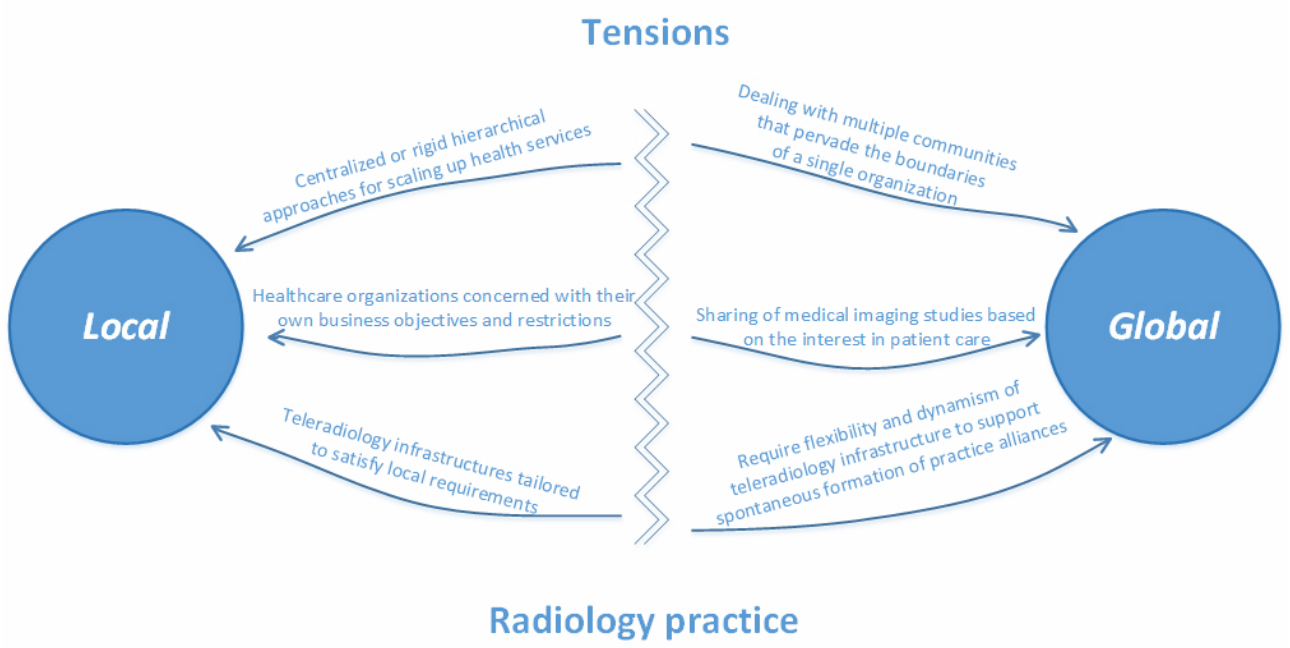
The big issues occur when health professionals need to access medical image studies outside the domain of a health care organization. Specifically, it is not easy for a physician to share an image study effectively for a second opinion with a colleague situated in a different location. By contrast, when all the actors are in the same (local) domain, it is easier to share the study through the radiology workflow of PACS/RIS. It is also easy to distribute finished reports to the referring physicians once they are in the same domain. In spite of advances in federated teleradiology solutions for integrating image sharing among health care organizations [16], they are complex to implement and such alliances involve business models. Ultimately, it would be of little interest for competing teleradiology companies to share medical imaging studies and other collaborative resources among themselves. On the other hand, the care of the patient should be of paramount interest for sharing medical imaging studies, so as to have an expert second opinion on a complex case. In such a situation, the consultant physician generally chooses the expert radiologist in a rather arbitrary manner, regardless of any kind of business alliances that may exist among health care organizations. Selection of the expert radiologist, as an illustration, may be based on the consultant physician's professional relationships, or on the recommendation of colleagues, or through reputation within a subspecialty. Essentially, there are many factors that affect the sharing of imaging studies for patient care, which require the flexibility and dynamism of teleradiology infrastructure to support the spontaneous formation of temporary or permanent practice alliances of health professionals and organizations. However, global health initiatives often adopt highly centralized or rigidly hierarchical approaches for scaling up health services, which are not fitting for the dynamic, unpredictable manners in which health services may expand and become sustainable [17]. In particular, teleradiology infrastructures of health care organizations tend to be tailored to satisfy local requirements. For instance, providing image interpretation or second opinion advice services to previously defined remote locations as part of a locally managed teleradiology service or, on the other hand, acting as a user of services provided by a specific teleradiology company. This leads to an emphasis on detailed initial planning and inflexible designs, which do not address the adaptive properties of dynamic pathways for expanding health services [17].

Despite the weaknesses of using email, in teleradiology it has become the most popular way to overcome this lack of flexibility and dynamism [18]. In its simplest form, the physician just collects the images of interest into some well-known format (eg, JPEG), packs and emails them to a remote expert, who will then review the images and reply with a report. In a more advanced manner, the registered DICOM MIME type [19] allows the transfer of imaging studies in DICOM standard format [20] using basic email transport mechanisms with additional encryption in accordance with OpenPGP [21]. Weisser et al [22] present the successful experience of integrating more than 60 health care organizations in Germany by transmitting DICOM imaging studies via email in a variety of teleradiology applications. Email has also been successfully used to send reports to women undergoing mammography screening in the United States [23]. Hence, what does the use of email in radiology suggest? It suggests email is a simple way to connect people and exchange medical imaging studies beyond the limited boundaries of local PACS networks. With email, it is easy to locate and connect people in order to exchange images or reports and collaborate asynchronously thanks to the simplicity, availability, connectivity, large number of users and low cost of email. Pianykh [18] claims that email radiology was "the first honest attempt to implement true teleradiology"; however, he also recognizes the drawbacks such as, poor image quality, the loss of metadata when images are converted from DICOM to common image formats (eg, JPEG), difficulties in dealing with large files, and a lack of any PACS/RIS integration and consolidated workflow.

It is manifest that there is a polarization between the local and global practice needs (Figure 1). In general, the local needs of the radiology practice are well afforded by the radiology workflow of the PACS/RIS in a health care organization, whereas there is a great impediment to support the global needs. This impediment is motivated by a radical difference between the local and global needs of the radiology practice. Health care organizations are often concerned with their own business objectives and restrictions, which influence the working practice of the local radiology community as well as the local PACS/RIS infrastructure. On the other hand, individuals, radiologists, patients and the many other stakeholders in the health care system are often members of multiple communities that pervade the boundaries of a single organization, interacting with one another through a web of complicated relationships influenced by communities of practice, neighborhoods and social networks [17]. Moreover, each practice community uses technologies differently, thus presenting different demands on their flexible standard requirements [7].

Such polarization establishes a tension between the local and global demands (ie, the demands that encompass the boundaries of a single organization) that denotes a lack of an information infrastructure for the radiology practice and this very infrastructure will occur only when this tension is resolved [2,7] by reconciling the local with the global. However, the emergence of such an information infrastructure is a long-term venture, which requires that it is considered not as something entirely transparent and ready to run or operate as something else, but as a social space of interrelations between people, organizations and technical components [2]. In fact, information infrastructures emerge not by emphasizing changes in the infrastructural components but from changes in the infrastructural relations, since information infrastructures are fundamentally a relational concept [2,7]. In this sense, to be social does not signify being a thing among other things, but a kind of association between things that are not themselves social, a movement that may fail to trace any new association and may fail to redesign any well-formed assemblage [24].

Figure 1. Polarization due to tensions between local and global demands for radiology practice.



Information Infrastructure for Radiology Practice as a Social Space

Overview

The information infrastructure for the practice of radiology means a social space of static and dynamic interactions where people, organizations and technical components are associated with activities and structures, forming a sociotechnical system. This social space may be a physical place, such as a radiologist's report room or a virtual space such as the radiology department, and it simultaneously offers material and immaterial support for social relations [25]. The material support for the practice of radiology, among others, includes physical rooms, furniture, medical imaging equipment, information technology (IT) devices, and networks. The immaterial support comprises business and clinical processes (activities) of the radiology department, organizational structure, roles and functions, IT and communication software (eg, PACS, RIS), among others. It is in this social space that people gather and interact with each other, with material and immaterial support. In addition, the social space for radiology practice is evolving and open.

The Social Space for Radiology Practice is Interactive

The interactions that occur in the social space for radiology practice can vary from rather static to highly dynamic. For instance, the relationship between a radiologist and an image modality tends to be rather static. Specifically, an expert radiologist in nuclear medicine (NM) tends to have a static relationship with the NM division of the radiology department in the sense that they belong to this division in the organizational structure for an indefinite time. In fact, such experts tend to work with the same set of imaging equipment located in specific

rooms of the NM division, use a familiar set of imaging manipulation tools and other information and communication technology (ICT) tools, carry out routine sets of clinical and administrative activities, and finally, usually cooperate daily with a well-known staff. In short, a health professional working in their practice in a social space establishes static relationships with material and immaterial support. In general, such relationships have a high inertia, change slowly over time, and are tied to a local context.

Other interactions are much more dynamic, such as the relationship between patients and the radiology department. Some patients go to the radiology department only once, while others make several visits during a short period of time, according to the condition of their health. Some visits are motivated by emergencies or the need for urgent examinations, while others are motivated by chronic diseases. One single patient may undergo tests with different imaging modalities, where images are acquired with the support of different kinds of equipment, clinical and administrative processes and various personnel, with the results being evaluated by different radiologists. Complex cases may require special care for patient preparation before the test, application of elaborate post-processing techniques on the acquired images, or the formation of medical boards to discuss findings. Certain demands, especially in emergencies or urgent examinations, may be highly dynamic and unpredictable. In sum, a patient in the social space of a radiology department establishes dynamic relationships with the physical environment, imaging modalities, software tools, clinical and administrative processes and personnel, among others. In general, the occurrence of such relationships is ephemeral and dependent on the patient's health and financial condition, but the outcomes may have a significant repercussion on the patient's life.

The Social Space for Radiology Practice is Evolving

With the aim of considering the social space as something that is evolving, it is essential to highlight the relational role played by the interactions between people, organizations, and technical components. This shifts the emphasis away from things and people as simply being causal factors during the performance of such practices [7], since interactions generate a chain of actions and reactions along complex pathways that influence the evolution of the social space in a variety of manners. For instance, the radiology department (a virtual place) interacts with the physical building of the hospital into which it is located. In this case, the physical building is continually adapting to satisfy the requirements of the radiology department while on the other hand, the physical restrictions of the building continuously affect the work practices in an iterative and interwoven mode. Similarly, the technical components are affected by the requirements of the radiology department, as well as by interactions with people and things, such as other technical components. However, this process is not only one-way, and people also place their interests in the technology [26]. As an illustration, it is possible for technology to change common activities within the radiological workflow in order to adapt to its interests, where radiologists, for example, have to learn a new method of entering diagnoses and other information onto a structured reporting system that has replaced the traditional use of free text editors. Conversely, technical components are modified and adapted to accommodate the demands of people and things, such as complying with a technical standard (eg, DICOM, Health Level Seven [27]) or changing the procedures for patient appointments, allowing them to also be booked via a mobile phone app. In short, the conventions of practice both shape and are shaped by information infrastructure [7]. Indeed, from the interplay of people, organizations and technical components in the social space there emerges a concurrent design and redesign of technology, individuals and work practices [26], forming an ever-evolving sociotechnical system.

The Social Space for Radiology Practice is Open

It is important to note that such interactions occur not only inside the local social space of a radiology department, but also within the external environment, since social spaces are inherently open, stretching outside their own location. This signifies that in spite of there being a boundary, a local social space is permeable to its external environment for the exchange of matter, energy and information. While on the one hand, such an exchange is essential for the social space of a radiology department to keep its internal organization and evolve, on the other, it also allows it to influence the (external) environment. However, being open does not denote being completely free because members of a social space have to adhere to rules, business and clinical processes, policies and principles, some self-determined and others determined from outside. For instance, not everyone is free to enter a radiology department to work as a radiologist. From an external perspective, there are legal requirements that the radiology department must observe regarding professional credentials in order to allow an applicant to work as a radiologist. However this is not enough. Based on its policies and principles, the radiology department itself should also have a particular interest in hiring a new radiologist, such as the need to expand the workforce to meet growing demands. Once hired, the radiologist becomes a member of a social space of work practice, in which membership signifies having to learn the rules [7], business and clinical processes, policies, and principles of the radiology department. Nonetheless, the radiologist, with their professional and personal history also has the potential to interfere and change the above mentioned items. Similarly, this also occurs with anyone who interacts with or within the social space of the radiology department, including other professionals and patients.

Openness can be helpful to deal with situations of emergency, particularly when they require a large effort in reading images that cannot be supported by a local radiology department alone. In this case, an open social space for radiology practice enables the temporary mobilization of a taskforce for reading images, with radiologists from the external environment being invited to join it. The transmission of DICOM image studies by email, like the experience related by Weisser et al [22], is an example of an open infrastructure for radiology practice that can help in such mobilization. In principle, any radiologist having an email account and a trustful digital certificate can join the effort, being able to receive studies to read and to send reports in a secure mode.

Other kinds of interactions with the external environment are more subtle, thanks to the openness of social spaces. Consider the case of clinical research carried out by a radiology department. The findings of such research need not be used only to improve the work practice locally, but also externally, since they are published through scientific conferences and journals and are kept in digital libraries alongside the findings of clinical research performed elsewhere, compounding a body of knowledge. Thus, local findings are potentially able to influence the work practices of other (local) radiology departments. However, it is worth noting that scientific societies, conferences, and journals mediate this flow of information between (local) radiology departments. In fact, they are also social spaces for radiology practice because people, organizations, and technical components interact when dealing with this body of knowledge. It is there that primary (eg, original papers) and secondary (eg, books, reviews, clinical guidelines, technical standards) research findings are discussed, peer-reviewed, and shared. As a social space, it is there that cooperation and competition takes place, recommendations are made, reputations are built, conflicts emerge, and consensus is reached. Furthermore, the body of knowledge collectively produced has an influence over the local work practices of radiology departments, where, in general, clinical research is conducted. This is feasible because the boundaries of social spaces are permeable to their external environment. Hence, at the same time, the environment is both intimate and foreign (it is part of a social space, yet it remains exterior to it), so that the intelligibility of a social space is encountered not only in the social space itself, but also in its relationship with the environment which is not simply of dependence: it is constitutive of it [28].

Essentially, the aforementioned social spaces, including their relationships with the environment, comprise an information infrastructure for radiology practice. In the past, the main form with which to share scientific information was through the means of physical delivery. Another way to share information was to attend scientific conferences or visit a radiology service. It may be stated that the sharing of scientific research information through the radiological social space was quite successful in the past. Nowadays, with advances in ICT, such sharing has largely improved, not only by the ease of access provided by digital networks and libraries, but also by the support provided by ICT in conducting scientific research itself. However, information published from clinical research has an important property: it is aggregated information. Usually, selected clinical and demographic data are collected from patients as part of clinical research protocol performed in a radiology department or within a group of radiology departments. Such information is usually processed by statistical methods, summarized and analyzed locally by those responsible for the research before the findings being published. Consequently, scientific papers present clinical information in a highly condensed and abstract manner. On one hand, this facilitates the spread of information throughout the radiological social space but on the other, it hinders access to private information from the patients who took part in the research. The outcome is that, taking into account the flow of scientific information in the radiological social space, there is no significant tension between local and global. The tension occurs when the clinical and demographic information that needs to be shared belongs to one identifiable individual.

Social Radiology Information Infrastructure Wrestles With the Inertia of an Installed Base

According to Star and Ruhleder [7], an information infrastructure is built on an installed base and wrestles with the inertia of this base, inheriting both its strengths and shortcomings. In the case of scientific information flowing into the radiological social space, ICTs, particularly the Internet and the Web, were introduced within an installed base that was mostly transposed to a virtual world, such as e-mails, e-documents, digital libraries, e-subscriptions, e-publishers, e-readers, and information systems to support scientific workflow. In short, the ease with which scientific information was commonly exchanged among local social spaces became even greater after the advent of ICTs, due to its growing pervasiveness. The ICT infrastructure to support radiology practices that deal with personal and identifiable clinical and demographic information was also built on an installed base. However, in this case, the radiological workflow, which deals with this kind of information, was traditionally confined to the boundaries of the radiology department. Before the advent of ICTs, the steps of the radiological workflow were performed within the physical space of the radiology department, where the generated medical images were made available on film and reports were written on paper. For situations requiring the opinion of a remote subspecialist radiologist, as in complex cases, the transmission of medical images over distance (ie, teleradiology) was not common due to technical difficulties of transmission, high costs, and poor image quality. In general, medical boards were formed to discuss complex cases with radiologists and other physicians from the local practice community of the radiology department or hospital. Concerns with the violation of patient confidentiality due to leaking sensitive clinical information also contributed to keeping the imaging studies within the confines of the radiology department that produced them.

Therefore, when ICTs were introduced into the radiology department to deal with personal and identifiable clinical and demographic information, they were used to support radiology practices that usually worked on a local basis. The examples include: (1) medical images generated on film were replaced by direct digital capture to produce a digital image available in DICOM standard format; (2) management of the physical films of medical images in the radiology department was replaced by PACS working according to DICOM requirements for image communication between individual components, such as imaging equipment, diagnostic and post-processing workstations, archive systems, and image distribution workplaces [10,11]; (3) the radiological workflow was mostly transposed to RIS/PACS, comprising software modules for creating orders, scheduling, reading, reporting, medical coding, recording services, and interfacing for billing systems, among others [11].

For cases in which teleradiology was required, it was only in the early 2000s that ICTs were ready for real clinical applications [10]. In general, teleradiology activities were supported by projects that created advanced infrastructures, although they were not sustainable, since they depended on short-term external resources that did not remain available after the end of the project [17]. In addition, such activities were conducted outside the radiology department routine, and did not complement it or become integrated. Thrall [8] reminds us that certain teleradiology efforts from the 1960s until the mid-1990s presented a relatively low performance as the cost of computers and data transmission were high, image quality was poor, and logistics were cumbersome. These efforts were unsustainable without external funding, and the clinical applicability in radiology work practices was very limited. By contrast, since the mid-1990s, particularly after the early 2000s, the evolution of ICTs provided a set of enhancements that enabled, in principle, an effective, sustainable use of teleradiology [8] as exemplified: (1) the availability of high-performance/low-cost personal workstations for image processing and display; (2) the availability of high-performance/low-cost storage and communications/computer networks like the Internet; (3) improvements in image compression algorithms and transmission techniques; (4) widespread use of PACS/DICOM by radiology departments.

While there is a inertia that confines the radiology work practices to the local social space of the radiology department, even after the arrival of ICTs the aforementioned enhancements bring very attractive opportunities to displace such work practices from the local to the global. In other words, they offer opportunities to disrupt the physical contiguity of the place where radiology work practices are usually performed. Fragments of the place need to be rearranged into a network in order to allow continuity of the work practices. In fact, there is still a movement in progress towards changing from a space of places to a space of flows, in terms of Castells' nomenclature [29]. The space of places (ie, the local social space of the radiology department) organizes experience and activity around the confines of a locality, while the space of flows electronically associates separate places in an interactive network that connects activities and people in distinct geographical contexts [30], that is, the global social space for radiology practice.

Current Teleradiology Models are not Information Infrastructures for Social Radiology

The movement against local inertia does not result in the social space vanishing from the radiology department. In fact, it results in its transformation by the new possibilities of organizing people, activities and structures. Today's common teleradiology models illustrate some of these possibilities [8,10,18]:

1. *Night-hawking/On-call/Off-hour reading*: these terms refer to providing on-call coverage for image interpretation, particularly during off-hours, when the availability of radiologists on examination sites is scarce. It is clear in the model that the radiology department has its own staff of radiologists, and on-call coverage is provided by designated members of the staff or by outsourced radiologists from a teleradiology company. In this last case, it is common to read images overnight in another country, taking advantage of a different time zone and lower costs.
2. *Regional PACS*: this model uses WAN to integrate local PACS or DICOM workstations from remote locations. It is typical for inter-hospital PACS for instance, when hospitals or health care centers have branches or satellite image centers, when they form business alliances, or when they are under the umbrella of a large public health system. It is a current solution for developing national and international radiology networks.
3. *Radiology outsourcing*: a model in which a teleradiology company takes care of the radiology service when interpretations are not available on site. In general, the hired company is in a cost-efficient location and may provide teleradiology equipment, image storage and technical support in addition to remote image interpretation by radiologists.

Although these models impact the inertia that confines the radiology work practices to the radiology department, they do not truly configure as an emergent information infrastructure for social radiology.

For the above listed items, in model (1), teleradiology does not substantially affect the usual work practices of the radiology department that takes advantage of it in two ways [8]. Firstly, by offering timely radiology coverage to referring physicians and patients, regardless of the availability of internal on-site staff, and secondly, by improving the usage of the workforce when the radiology department has its own 24-hour staff coverage, taking advantage of this to offer image interpretation services to third parties. In short, teleradiology model (1) is used as a convenience, to enhance the usage of the local workforce, or to maintain a reasonable work life for local staff.

Teleradiology model (2) lacks the flexibility required by an information infrastructure. The integration of several local PACS is driven by the business concerns of a relatively small group of hospitals or health care centers, not by the true concern for sharing medical images in general. The difficulties of sharing images outside the regional PACS domain remain, as much of the conventional PACS inertia is inherited [18]. For instance, a common approach is to have a VPN via WAN connecting branches and satellite image centers to the central PACS of a main hospital. In this case, a regional PACS is essentially a conventional, but huge PACS [18]. For the case of business alliances involving few hospitals, customized solutions to integrate their PACS are common, but there are problems of interoperability. The alignment of business interests among the participants of a regional PACS, on the other hand, facilitates sharing efforts and cooperation because a trust relationship is a priori established. In spite of this, the regional PACS is merely an integration of local PACS among organizations with a great interest in such sharing, offering nothing new in relation to conventional PACS. High inertia to planning and maintaining regional PACS hampers the flexibility and dynamism required to support the spontaneous formation of temporary or permanent

practice alliances of health professionals and organizations motivated by the interest in patient care. In sum, teleradiology model (2) does not significantly affect local work practices, nor does it facilitate the exploration of new possibilities for radiology work practices outside the domain of regional PACS members.

Finally, the main problem in model (3) is the dependence on a single company that, in general, provides an ad hoc infrastructure for teleradiology suited for making easy, fast, and cheap connections with clients. The design of such solutions tends to be inflexible, and does not address the complexities of interoperability, because teleradiology companies generally have no interest in sharing images with external entities. Even so, this model of teleradiology may impact the local social space of radiology work practice because it displaces the radiologists from the point of patient care to another location. Motivations for this displacement (ie, the absence of interpretations on site) are of a logistical and economic nature. Logistical motivation is a classic case for employing teleradiology: remote areas (eg, rural, difficult to reach and possibly with only non-radiologist physicians) using a remote service for image interpretation by radiologists, including emergency situations and for second opinions. The impact in this case is positive. The economic motivation, on the other hand, aims to reduce the costs of maintaining an onsite team of radiologists. Local staff is replaced by remote radiologists hired by teleradiology companies situated in a cost-efficient location. One criticism of this last motivation is that it leads radiology work practice towards commoditization (assembly-line approach), as teleradiology companies and hospitals seek to maximize financial gain, without due concern for consultative skills, the necessary assessment or quality control provided by radiologists [15].

Reconciling the Global With the Local

Essentially, in such common teleradiology models, work practice from the confines of locality is not truly reconciled into a space of flows to form a global social space for radiology. While they fragment the physical contiguity of the place where radiology work practice is performed, such fragments somehow remain close, due to local inertia. As a result, the impediment of sharing medical imaging studies with other localities remains high, since tension between the local and global persists, thus hindering the emergence of social radiology as an information infrastructure.

Observing from the perspective of Marc Berg's Rationalizing Medicine [31], there was a convergence between technological tools and radiology practice when ICTs were introduced in radiology departments. The very creation of the DICOM standard, PACS and RIS as well as new or reshaped radiology practices, commonly found today on local radiology departments, was "not pre-given, but emerged in and through the development and intertwining of networks" [31], involving health professionals, technicians, patients, organizations, among other stakeholders. Such convergence, on the other hand, was not observed in the common teleradiology models presented beforehand in order to signal a seamless integration of local and global into an emerging information infrastructure for radiology practice. This suggests that such models are some of the "many loose ends" that confronts processes of convergence [31], still in progress towards social radiology information infrastructure.

An effective information infrastructure for radiology practice should facilitate social interaction regardless of any kind of business alliances among health care organizations. Here, reconciliation between local and global signifies teleradiology as an integral part of PACS, being the notion of (local) PACS and (global) teleradiology transparent, with digital imaging without the constraints of distance becoming a true radiology standard [18]. In fact, this facilitation will be reached insofar as the information infrastructure is shared, open, heterogeneous, secure, and evolving, forming a sociotechnical system of information technology.

A shared information infrastructure means considering it as a public good [2], and not belonging to a single company or organization, but shared across multiple communities in many, unexpected ways [1]. An open information infrastructure means that it has permeable boundaries, which allow interactions with an external environment in intricate, unexpected manners and contexts. In fact, the boundaries are not clear enough to distinguish those that may use the information infrastructure and those that may not, nor those that may design the information infrastructure and those that may not [1]. Heterogeneity reflects the great social and technical diversity afforded by an open, shared information infrastructure, able to include a growing number of entities such as user communities, operators, governance and standardization bodies, and design communities [1,4]. A secure information infrastructure signifies the capability of establishing trust among the entities of which it is composed, in consideration of legal and ethical issues such as patient privacy, confidentiality, integrity and ownership of clinical data, licensure, accreditation and liability of health professionals and organizations [16]. The experience of the DICOM email in Germany [22] is an example of an open and loosely coupled infrastructure for teleradiology that addressed such security issues. Finally, an evolving information infrastructure signifies considering it as emerging from the continuous interplay of people, organizations and technical components in a concurrent process of design and redesign [26].

In point of fact, social radiology as an information infrastructure (Figure 2) is a social space of static and dynamic interactions for radiology practice where people, organizations and technical components are associated to activities and structures forming a sociotechnical system of information technology that is shared, open, heterogeneous, secure, and evolving. It may enable the reorganization of radiology work practice into cyberspace (space of flows) to form a global social space for radiology that surpasses current teleradiology models. As a sociotechnical system of information technology, it is a basis for social computing that may provide value way beyond that offered by purely IT systems, since user-generated content is exploitable not only by the users, but by the information infrastructure itself [32].

The information infrastructure may be of value by producing faster results due to multiplying effort [32]. For instance, the

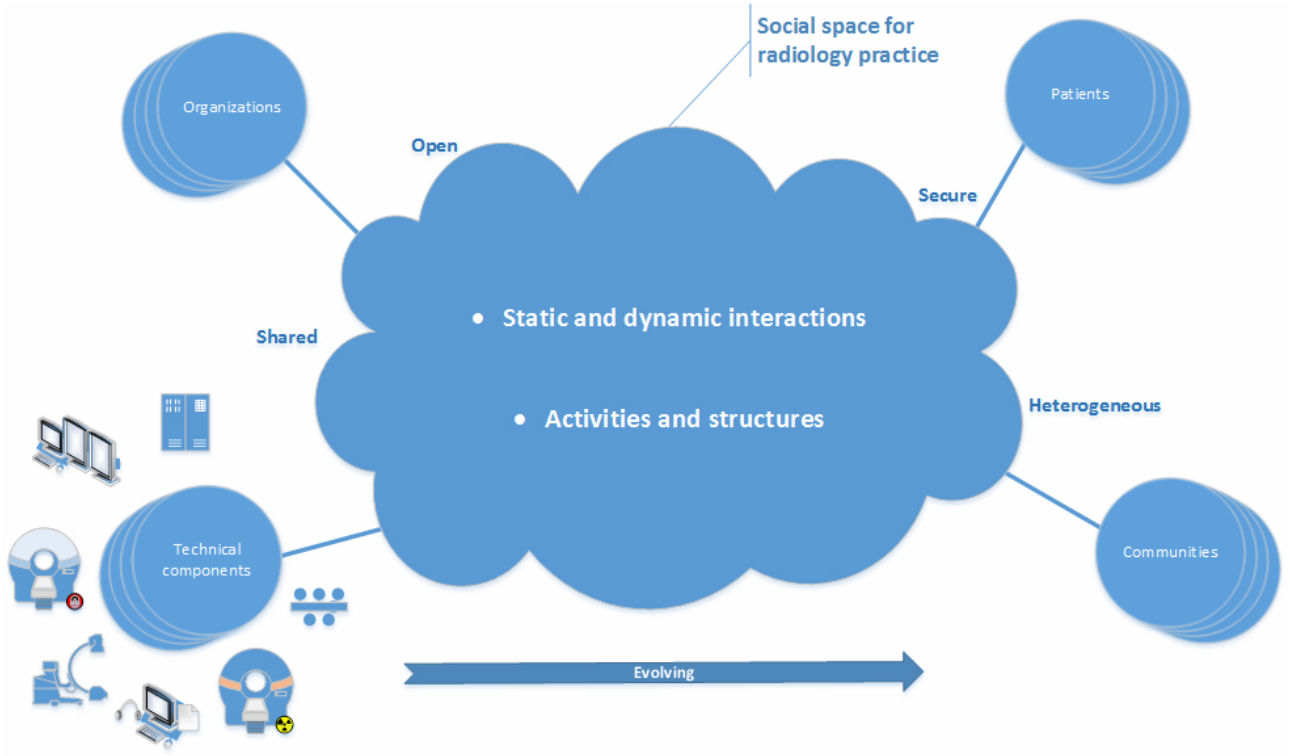
information infrastructure may facilitate the spontaneous formation of small groups of subspecialist radiologists to provide expert consultations [15]. The agglutination of such groups to form larger groups may additionally provide 24/7 coverage for several small organizations and thus, the responsibility for off-hour emergency examinations, shared and spread over a large number of people, is able to enhance productivity of local workforce usage while maintaining a reasonable work load [33]. By being open and shared, a social radiology information infrastructure empowers radiologists to come together to provide professional services without the need of a teleradiology company acting as a broker.

The information infrastructure may also be of value by producing high quality results because it enables the integration of knowledge from multiple professionals with diverse expertise [32]. The existence of networks of groups of subspecialists favors the development of a culture of reciprocity in asking colleagues for advice and second opinions [33]. Indeed, it favors the creation of new models for assessing the quality of the radiologist's work and for peer review [34], as they have been challenged by referring physicians and health care organizations to demonstrate the quality and accuracy of their interpretations more objectively [8]. This may result in solutions that tackle the problem of commoditization in radiology by enhancing the work of the radiologists while considering patient benefit essential [15]. In addition, the pursuing for quality favors groups of subspecialists to create their own culture and standards for reading images. As such, the different cultures for reading images present in local radiology departments can also happen in the cyberspace because an open, heterogeneous and flexible information infrastructure enables such diversity.

Another way in which the information infrastructure may be of value is by producing results that are perceived as more legitimate because they represent a community [32]. For instance, a group of radiologists that provides expertise consultation is part of a community that may assess the results produced by the members of the group using some kind of socially constructed recommender system. Such a system is socially constructed, as it reflects a congruence between the behavior of the members of the group of radiologists (legitimate entity) and the (assumedly) shared beliefs of the community; therefore legitimacy depends on a collective audience, although it is independent of particular observers [35]. In this sense, legitimacy is seen as a social judgment of acceptance, suitability and desirability [36]. A basic premise to this is that the individuals of the community have an identity to enable interaction and communication, and association with the information produced [32], as legitimacy is dependent on an individual's history of events [35]. In fact, involvement of the community in building legitimacy for the radiology practice is essential so as to reinforce the growth and sustainability of the very community in the radiology social space, because legitimacy is an important factor for attracting resources from the external environment to maintain such growth and sustainability [36]. It is also important to empower patients in the relationship with radiology practice, either by providing information to support decision-making, such as choosing an expert for a second opinion, or by offering the possibility to evaluate the actions of a professional.

Finally, the information infrastructure may be of value by executing tasks that require exclusively human abilities, beyond the capacity of purely IT systems [32]. This is the case of interpreting medical images, a complex task that IT systems generally cannot perform alone. The task involves the process of image perception to identify abnormal patterns, followed by characterization and interpretation of those patterns [37], which depend heavily on empirical knowledge, memory, intuition, and diligence of the radiologist [38]. Despite this, computer-aided diagnosis (CAD) can be a helpful tool to support the radiologists in decision making, particularly in the process of identifying abnormal patterns. For example, studies have demonstrated improved diagnostic sensitivity with the use of CAD for assessing breast nodules, although with increased false-positive results [37]. This CAD could be useful in large-scale breast screening programs to pre-select imaging studies where possible breast nodules were detected, to distribute them among a taskforce of radiologists from groups of subspecialists who provide expertise consultation to the social radiology information infrastructure. The radiologists would use the CAD results as a "double-check" and in such a case, with increased false-positive results, this may help to reduce inter-observer variability among radiologists [37]. It is noteworthy that the final decision lies with the radiologists, providing additional value due to the synergistic effect of combining the radiologist's skills and the IT system's capability [39].

Figure 2. Social radiology as an information infrastructure.



Towards Social Radiology as an Information Infrastructure

This article explores the concept of social radiology as an information infrastructure, showing that the persistent tension between the local and global demands for radiology practice is an impediment for the emergence of such an information infrastructure. Tension persists due to the inertia of locally installed bases in radiology departments, for which common teleradiology models are not truly capable of reorganizing as a global social space for radiology practice. Reconciliation between local and global will facilitate the sharing of medical imaging studies beyond local domains, allowing the spontaneous formation of temporary or permanent practice alliances of (groups of) health professionals and organizations in a flexible and dynamic manner. With this reconciliation, the conceptual boundaries between (local) PACS and (global) teleradiology will vanish, signaling the emergence of social radiology as an information infrastructure.

The challenge is how to induce a movement to build social radiology as an information infrastructure, considering that it involves addressing a variety of issues, which are beyond the local and global tension examined in this article, for example, the tension between social and technical demands for radiology practice that arises among members of users and design communities, governance and standardization bodies, and health care organizations. More specifically, this tension is present in the sociotechnical process of developing information infrastructure standards that increases irreversibility in the use of technologies (eg, DICOM) while being open to further change and supporting flexibility of use [40]. This sociotechnical tension is also present in the relationship between user and open source software communities with traditional companies of medical imaging software [41]. Individual versus community demands are also a source of tension [2], and must be addressed in the move to build social radiology as an information infrastructure as well as security questions concerning the establishment of trust among the sociotechnical entities comprising the information infrastructure.

In the face of all these issues, recent advances in information infrastructure research [1-7], particularly in information infrastructure design theories [42], provide a promising way towards solutions for building social radiology as an information infrastructure. The design theory for dynamic complexity in information infrastructure [1] is a normative design theory systematized from empirical descriptions of the evolution of information infrastructures that tackles dynamic complexity in the design for information infrastructures, defined as a sociotechnical system of information technology. According to the proposed theory, information infrastructures have evolutionary dynamics that are nonlinear, path dependent and influenced by unbounded user and designer learning, as well as by network effects. In addition, information infrastructures are regulated by emergent, distributed, episodic forms of control. Therefore, information infrastructure design theory is aligned with a new view of health systems (such as a social radiology information infrastructure) as complex adaptive systems [17,43]. However, more research is needed regarding the application of design theories aimed at building information infrastructures for health systems, especially, social radiology.

Acknowledgments

This work was supported by the National Institute of Science and Technology for Software Engineering (INES), funded by CNPq, grants 573964/2008-4. I would like to thank Stuart Anderson, University of Edinburgh, for helpful input in this work.

Conflicts of Interest

None declared.

References

1. Hanseth O, Lyytinen K. Design theory for dynamic complexity in information infrastructures: the case of building internet. J Inf Technol 2010 Mar;25(1):1-19. [doi: 10.1057/jit.2009.19]
2. Bowker GC, Baker K, Millerand F, Ribes D. Toward information infrastructure studies: ways of knowing in a networked environment. In: Hunsinger J, Klastrup L, Allen M, editors. International handbook of internet research. Dordrecht: Springer; 2010:97-117.
3. Ure J, Procter R, Lin Y, Hartswood M, Anderson S, Lloyd S, et al. The development of data infrastructures for e-health: a socio-technical perspective. J Assoc Inf Syst 2009;10(5):415-429 [FREE Full text]
4. Edwards PN, Jackson SJ, Bowker GC, Knobel CP. Understanding Infrastructure: Dynamics, Tensions, and Design. 2007 Jan. Report of a Workshop on "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures" URL: http://deepblue.lib.umich.edu/bitstream/handle/2027.42/49353/UnderstandingInfrastructure2007.pdf?sequence=3 [accessed 2014-09-22] [WebCite Cache ID 6QapOdcgL]
5. Edwards PN, Jackson SJ, Chalmers MK, Bowker GC, Borgman CL, Ribes D, et al. Knowledge Infrastructures: Intellectual Frameworks and Research Challenges. 2013 May. Report of a workshop sponsored by the National Science Foundation and the Sloan Foundation URL: http://deepblue.lib.umich.edu/bitstream/handle/2027.42/97552/Edwards_etal_2013_Knowledge_Infrastructures.pdf?sequence=3 [accessed 2014-08-11] [WebCite Cache ID 6RkrIF1V1]
6. Hanseth O, Aanestad M. Design as bootstrapping. On the evolution of ICT networks in health care. Methods Inf Med 2003;42(4):385-391. [doi: 10.1267/METH03040385] [Medline: 14534638]
7. Star SL, Ruhleder K. Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. Information Systems Research 1996 Mar;7(1):111-134. [doi: 10.1287/isre.7.1.111]
8. Thrall JH. Teleradiology. Part I. History and clinical applications. Radiology 2007 Jun;243(3):613-617. [doi: 10.1148/radiol.2433070350] [Medline: 17517922]
9. Levin DC, Rao VM, Parker L, Frangos AJ, Sunshine JH. Bending the curve: the recent marked slowdown in growth of noninvasive diagnostic imaging. AJR Am J Roentgenol 2011 Jan;196(1):W25-W29. [doi: 10.2214/AJR.10.4835] [Medline: 21178027]
10. Huang HK. PACS and Imaging Informatics: Basic Principles and Applications. New Jersey: Wiley-Blackwell; 2010.
11. Zapf C, Bermann A, Sunderbrink D. PACS and RIS. In: Kramme R, Hoffmann KP, Pozos RS, editors. Springer handbook of medical technology. Berlin, Heidelberg: Springer; 2011:1199-1208.
12. Silva LA, Costa C, Oliveira JL. DICOM relay over the cloud. Int J Comput Assist Radiol Surg 2013 May;8(3):323-333. [doi: 10.1007/s11548-012-0785-3] [Medline: 22875554]
13. Figueiredo JFM, Motta GHMB. SocialRAD: an infrastructure for a secure, cooperative, asynchronous teleradiology system. Studies in Health Technology and Informatics 2013;192:778-782. [doi: 10.3233/978-1-61499-289-9-778]
14. Langer SG, Persons K, Erickson BJ, Blezek D. Towards a more cloud-friendly medical imaging applications architecture: a modest proposal. J Digit Imaging 2013 Feb;26(1):58-64 [FREE Full text] [doi: 10.1007/s10278-012-9545-8] [Medline: 23135215]
15. Borgstede JP. Radiology: commodity or specialty. Radiology 2008 Jun;247(3):613-616. [doi: 10.1148/radiol.2473072159] [Medline: 18487531]
16. Ribeiro LS, Costa C, Oliveira JL. Current trends in archiving and trasmission of medical images. In: Erondu OF, editor. Medical Imaging. Rijeka, Croatia: InTech; 2011:89-106.
17. Paina L, Peters DH. Understanding pathways for scaling up health services through the lens of complex adaptive systems. Health Policy Plan 2012 Aug;27(5):365-373 [FREE Full text] [doi: 10.1093/heapol/czr054] [Medline: 21821667]
18. Pianykh OS. DICOM and teleradiology. In: Digital imaging and communications in medicine (DICOM). Berlin: Springer; 2012:281-317.
19. DICOM Standards Committee. DICOM supplement 54 (DICOM- E-mail). 2000. URL: http://medical.nema.org/DICOM/supps/sup54_pc.pdf [accessed 2014-06-25] [WebCite Cache ID 6QapkNBQV]
20. DICOM Standards Committee (2011) Digital Imaging and Communication in Medicine (DICOM)-Part 1-20. Rosslyn, VA: National Electrical Manufacturers Association URL: http://medical.nema.org/standard.html [accessed 2014-06-25] [WebCite Cache ID 6QapxiFnO]
21. Network Working Group. RFC4880: OpenPGP message format. 2007. URL: http://www.ietf.org/rfc/rfc4880.txt [accessed 2014-06-25] [WebCite Cache ID 6QaqDmFtG]

22. Weisser G, Engelmann U, Ruggiero S, Runa A, Schröter A, Baur S, et al. Teleradiology applications with DICOM-e-mail. Eur Radiol 2007 May;17(5):1331-1340. [doi: 10.1007/s00330-006-0450-8] [Medline: 17031452]
23. Hall FM. The radiology report of the future. Radiology 2009 May;251(2):313-316. [doi: 10.1148/radiol.2512090177] [Medline: 19401567]
24. Latour B. Reassembling the social: an introduction to actor-network-theory. Oxford: Oxford University Press; 2005.
25. Lefebvre H. State, space, world: selected essays. Minneapolis, Minn: University of Minnesota Press; 2009.
26. Aanestad M, Hanseth O. Implementing open network technologies in complex work practices: a case from telemedicine. In: Baskerville R, Stage J, DeGross JI, editors. Organizational and Social Perspectives on Information Technology. Aalborg: Springer; 2000:355-369.
27. Health Level Seven International. 2014. URL: http://www.hl7.org/ [accessed 2014-06-25] [WebCite Cache ID 6Qar0cAte]
28. Morin E. On Complexity. Cresskill, New Jersey: Hampton Press; 2008.
29. Castells M. An introduction to the information age. City 1997 May;2(7):6-16. [doi: 10.1080/13604819708900050]
30. Castells M. Space of flows, space of places: materials for a theory of urbanism in the information age. In: Graham S, editor. The cybercities reader. London: Routledge; 2003:82-93.
31. Berg M. Rationalizing medical work: decision-support techniques and medical practices. Cambridge, Mass: MIT Press; 1997.
32. Erickson T. Social computing. In: Soegaard M, Dam RF, editors. The encyclopedia of human-computer interaction. Aarhus, Denmark: The Interaction Design Foundation; 2013.
33. Thrall JH. Teleradiology. Part II. Limitations, risks, and opportunities. Radiology 2007 Aug;244(2):325-328. [doi: 10.1148/radiol.2442070676] [Medline: 17641358]
34. Kaewlai R, Abujudeh H. Peer review in clinical radiology practice. AJR Am J Roentgenol 2012 Aug;199(2):W158-W162. [doi: 10.2214/AJR.11.8143] [Medline: 22826416]
35. Suchman MC. MANAGING LEGITIMACY: STRATEGIC AND INSTITUTIONAL APPROACHES. Academy of Management Review 1995 Jul 01;20(3):571-610. [doi: 10.5465/AMR.1995.9508080331]
36. Zimmerman MA, Zeitz GJ. BEYOND SURVIVAL: ACHIEVING NEW VENTURE GROWTH BY BUILDING LEGITIMACY. Academy of Management Review 2002 Jul 01;27(3):414-431. [doi: 10.5465/AMR.2002.7389921]
37. Li KC, Marcovici P, Phelps A, Potter C, Tillack A, Tomich J, et al. Digitization of medicine: how radiology can take advantage of the digital revolution. Acad Radiol 2013 Dec;20(12):1479-1494. [doi: 10.1016/j.acra.2013.09.008] [Medline: 24200474]
38. Tourassi GD. Journey toward computer-aided diagnosis: role of image texture analysis. Radiology 1999 Nov;213(2):317-320. [doi: 10.1148/radiology.213.2.r99nv49317] [Medline: 10551208]
39. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imag Grap 2007;31:198-211. [doi: 10.1016/j.compmedimag.2007.02.002]
40. Hanseth O, Monteiro E, Hatling M. Developing Information Infrastructure: The Tension Between Standardization and Flexibility. Science, Technology & Human Values 1996 Oct 01;21(4):407-426. [doi: 10.1177/016224399602100402]
41. Ratib O, Rosset A, Heuberger J. Open Source software and social networks: disruptive alternatives for medical imaging. Eur J Radiol 2011 May;78(2):259-265. [doi: 10.1016/j.ejrad.2010.05.004] [Medline: 21444166]
42. Gregor S, Jones D. The anatomy of a design theory. J Assoc Inf Syst 2007;8:312-335 [FREE Full text]
43. Sturmberg JP, Martin CM. Handbook of Systems and Complexity in Health. New York, NY: Springer; 2013.

Abbreviations

CAD: computer-aided diagnosis
CNPq: National Counsel of Technological and Scientific Development
DICOM: Digital Imaging and Communications in Medicine
ICT: information and communication technology
INES: National Institute of Science and Technology for Software Engineering
IT: information technology
JPEG: Joint Photographic Experts Group
MIME: multipurpose internet mail extensions
NM: nuclear medicine
OpenPGP: open pretty good privacy
PACS: picture archive and communication systems
RIS: radiology information systems
VPN: virtual private networks
WAN: wide area network

Edited by G Eysenbach; submitted 25.06.14; peer-reviewed by G Bowker, A Dheer; comments to author 24.07.14; revised version received 14.08.14; accepted 31.08.14; published 03.10.14.

Please cite as:
Motta GHMB
Towards Social Radiology as an Information Infrastructure: Reconciling the Local With the Global
JMIR Med Inform 2014;2(2):e27
URL: http://medinform.jmir.org/2014/2/e27/
doi:10.2196/medinform.3648
PMID:25600710

Original Paper

Enabling Online Studies of Conceptual Relationships Between Medical Terms: Developing an Efficient Web Platform

Aaron Albin[1,2], MS; Xiaonan Ji[1,2], MS; Tara B Borlawsky[1], MS; Zhan Ye[3], PhD; Simon Lin[4], MD; Philip RO Payne[1], PhD; Kun Huang[1], PhD; Yang Xiang[1], PhD

[1]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, United States

[2]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, United States

[3]Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI, United States

[4]The Research Institute, Nationwide Children's Hospital, Columbus, OH, United States

Corresponding Author:
Yang Xiang, PhD
Department of Biomedical Informatics
The Ohio State University
250 Lincoln Tower
1800 Canon Drive
Columbus, OH, 43210
United States
Phone: 1 614 366 5066
Fax: 1 614 688 6600
Email: yxiang@bmi.osu.edu

Abstract

Background: The Unified Medical Language System (UMLS) contains many important ontologies in which terms are connected by semantic relations. For many studies on the relationships between biomedical concepts, the use of transitively associated information from ontologies and the UMLS has been shown to be effective. Although there are a few tools and methods available for extracting transitive relationships from the UMLS, they usually have major restrictions on the length of transitive relations or on the number of data sources.

Objective: Our goal was to design an efficient online platform that enables efficient studies on the conceptual relationships between any medical terms.

Methods: To overcome the restrictions of available methods and to facilitate studies on the conceptual relationships between medical terms, we developed a Web platform, onGrid, that supports efficient transitive queries and conceptual relationship studies using the UMLS. This framework uses the latest technique in converting natural language queries into UMLS concepts, performs efficient transitive queries, and visualizes the result paths. It also dynamically builds a relationship matrix for two sets of input biomedical terms. We are thus able to perform effective studies on conceptual relationships between medical terms based on their relationship matrix.

Results: The advantage of onGrid is that it can be applied to study any two sets of biomedical concept relations and the relations within one set of biomedical concepts. We use onGrid to study the disease-disease relationships in the Online Mendelian Inheritance in Man (OMIM). By crossvalidating our results with an external database, the Comparative Toxicogenomics Database (CTD), we demonstrated that onGrid is effective for the study of conceptual relationships between medical terms.

Conclusions: onGrid is an efficient tool for querying the UMLS for transitive relations, studying the relationship between medical terms, and generating hypotheses.

(JMIR Med Inform 2014;2(2):e23) doi:10.2196/medinform.3387

KEYWORDS

UMLS; ontology; conceptual relationships

Introduction

Since Swanson's discovery of the connection between fish oil and Raynaud's syndrome via blood viscosity [1], transitive associations have been important sources of hypothesis generation in biomedical science. In Swanson's paradigm, an association between concepts A and C may be possible if both are related to a third concept, B. A number of discoveries and hypotheses have been made under this model. For instance, Hristovski et al proposed literature-based discovery to search disease candidate genes [2], to investigate drug mechanisms [3], and to identify novel therapeutic approaches [4]. As another example, Petric et al used this model to study autism by literature mining and found the connection between autism and calcineurin [5]. With the Unified Medical Language System (UMLS), such transitive association studies are becoming more efficient and powerful in generating novel hypotheses.

In biomedical science, the UMLS [6] is the largest thesaurus widely used in various applications. It is a collection of more than 160 source vocabularies (version 2012AA). The UMLS consists of three parts: the Metathesaurus, Semantic network, and Specialist lexicon. The Metathesaurus is the main body of the UMLS and has over 2 million concepts, each with a concept unique identifier (CUI), and over 15 million links (associations) between pairs of CUIs. The UMLS Terminology Services (UTS), hosted by the National Library of Medicine, provides an online query tool for these concepts under its Metathesaurus browser. To make use of the rich information contained in the UMLS, the interactive biomedical discovery support system (BITOLA) developed by Hristovski et al [2,7] supports the input of UMLS CUIs, concept, semantic types, and chromosome locations, in searching for hypothetic relations such as disease candidate genes.

BITOLA is based on Swanson's one transitive relationship model. It is quite natural to ask if multiple transitive relationships will generate more rich hypotheses. Wilkowski et al [8] showed that by extracting paths from a graph modeling the concept relations, it is possible to extend this one transitive relationship model to a multiple-transitive relationship model for novel hypothesis discovery. For the UMLS, if we consider each CUI as a vertex, and links connecting two CUIs as an edge, we obtain a graph modeling the UMLS. The transitively associated queries on the UMLS can be regarded as queries on the UMLS graph. In fact, a number of works [9-14] have successfully used multiple-transitive relationships in the UMLS to study the closeness between two medical concepts. However, these works have two major limitations.

First, similar to [8], they rely on ad-hoc path search algorithms, such as Depth-First Search (DFS), which limit their searching ability on very large graphs. This is because the running time of DFS or similar ad-hoc search algorithms is proportional to the size of the graph being searched. As a result, it is not efficient to perform a large number of searches on a large graph using these algorithms. Thus, these works put major limitations on their search ranges, such as within a very small number of data sources in the UMLS, or very short search paths (eg, no more than 5 concepts in a path in [11]), to reduce the search space and thus to reduce the search time. Second, they generally rely on the distance to determine the closeness between two concepts. Since the distance between two concepts is determined by the shortest transitive relationship(s) and does not take into account other non-shortest transitive relationships, a false shortest transitive relationship may nullify the whole hypothesis. Given this observation, we conclude that this is not as reliable as a measurement on a large collection of paths. In fact, the effective measurement of relationship between two concepts in [2] and [15] can be viewed as a measurement on a collection of very short paths.

To overcome the two limitations, we developed a k-neighborhood Decentralization Labeling Scheme (kDLS) to efficiently index the UMLS [16]. kDLS supports efficient path/distance queries on the whole UMLS, as well as a measurement on the closeness between any two UMLS concepts by a collection of paths found between them. Efficiently querying such a large graph is a significant challenge for the graph database community. In fact, even the very recent graph indexing scheme [17] does not demonstrate the ability to efficiently answer distance queries on graphs with similar size and density. kDLS utilizes the power-law property of the UMLS for designing the indexing algorithm and turns out to be very effective in indexing the UMLS for both answering graph queries and discovering knowledge. Explained briefly, the indexing algorithm of kDLS iteratively removes a high degree vertex from the UMLS graph and broadcasts its information to the remaining vertices in the k neighborhood of the removed vertex. When the indexing ends, each vertex has a list of records that is considered its label. By comparing the labels of two vertices, it is possible to find a collection of paths (including but not limited to shortest paths) between the two vertices. We have proven that kDLS is guaranteed to find at least one shortest path if the two vertices are within k hops on the UMLS graph. On average, the number of paths discovered by kDLS is much larger than by the DFS or the Breadth-First Search (BFS), as we have shown previously [16]. Subsequently, the measurement between two concepts is based on the number of paths discovered as well as their lengths. kDLS has demonstrated its power in medical concept coreference resolution in clinical text [18].

However, kDLS has several major disadvantages: (1) it does not take into account the semantic networks in the UMLS ontologies, (2) it does not accept natural language–based queries, and only accepts queries on UMLS CUIs, and (3) it is time costly and difficult to configure and use kDLS for one study, regardless of the size of the study. To address these disadvantages, we developed an efficient online conceptual study platform using Graph indexing, onGrid, to study the conceptual relationships between biomedical terms.

Methods

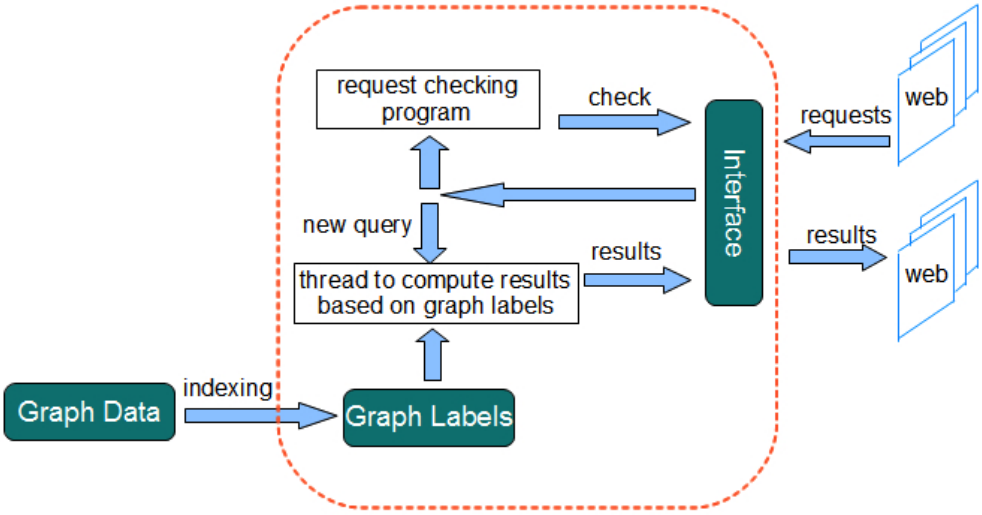
System Framework

The cost to load the kDLS index is a major limitation of kDLS. Typically, it requires more than 20GB of memory [16] and takes several hours to load the kDLS index into memory before it can be used to efficiently answer queries and output discovered

results. To provide an efficient solution for studies on conceptual relationships between medical terms, we developed onGrid, an online conceptual study platform using Graph indexing. onGrid provides a user-friendly Web interface to accept natural language-based queries and convert the queries to index-based searching on the UMLS and is expected to support future graph index engines on the UMLS. In addition, we proposed a new indexing method for onGrid that takes into account the concept semantic types, and our study on conceptual disease relationships demonstrated the advantages of the proposed indexing method over the original kDLS indexing method.

The general framework of onGrid consists of two parts: the client side, which was implemented in JavaScript and PHP (Hypertext Preprocessor), and the server side, which was implemented in C++, a general purpose programming language. The client side receives query requests from users and transmits them to the server, which then executes the query requests and sends the results back to the client. This design pushes the light and fast pre-computation and post-computation tasks to the client side, which has limited resources, and the computing-intensive tasks to the server. To realize this, the server side program first loads the graph index into memory and iteratively checks for new requests from the client side. Once a new query request is received, the server side program dispatches a new thread to handle the request by using the loaded graph index. When the thread completes the request, it saves the results to be retrieved from the client. We use a MySQL database as the interface to facilitate the communication between the client side and the server side. All requests and results are posted to the database, which is regularly checked by both the server and client side programs. The flowchart of the system framework is illustrated in Figure 1.

Figure 1. Flowchart of the onGrid framework.



Web-Based Natural Language Processing

To enable natural language-based queries on the UMLS, we developed LDPMap [19], a layered dynamic programming approach that maps a biomedical concept to a UMLS concept. Since UMLS is very comprehensive, nearly all medical concepts can find their corresponding part in the UMLS. Our study shows that LDPMap is an effective tool for mapping a biomedical concept to a UMLS concept. In this work, we integrate LDPMap into onGrid such that biomedical terms in a query will be mapped to UMLS concepts before the query is executed. To avoid mapping errors, the system will automatically provide a list of mapped UMLS terms with CUIs in order of relevance for querying the relationship between two medical terms. Users can accurately select the terms for further querying.

Network Visualization

Querying for relationships between two concepts returns a collection of paths between two query concepts. To provide users intuition on the path query results, onGrid visualizes the shortest paths among these paths. Visualizing all paths may not be feasible because the path query results often contain thousands of paths or more, which are hardly discernible considering the visual clutter. On visualizing the shortest paths between two vertices u and v, we organize all vertices that have the same distance to vertex u (or v) into a set S_k where $S_k = \cup_{p \in P'(u,v)}\{x|x \in p, distance(x,u)=k\}$ ($P'(u,v)$ is the set of shortest paths among the collection of paths between u and v. All vertices in a set S_k will be visualized on a line perpendicular to the line connecting u and v. In this way, we are not only able to observe paths connecting two vertices but also observe shared vertices and edges among those paths.

Concept Similarity Measurement

To measure the closeness between two concepts, onGrid takes into account the semantic type of each concept (vertex). UMLS (version 2012AA) has a total of 133 concept semantic types such as “Event”, “Disease or Syndrome”, etc. They are organized in a directed acyclic graph known as the UMLS concept semantic network. The semantic types closer to the root level are more abstract than those closer to the leaf level. Abstract semantic types are more likely to be related to a large number of concepts, and therefore we consider such relationships weak. To put more emphasis on concrete concepts in a path, the closeness between two concepts are measured by:

$R(u,v)=\sum_{p\in P(u,v)}\prod_{x\in p}g(x)$

where $P(u,v)$ is the collection of all paths between u and v discovered by kDLS, excluding paths with lengths equal to 1. $g(x)$ is the semantic function on vertex (concept) x. In the onGrid implementation, we let $g(x) = 1/h$ where h is the reverse topological level of vertex x. All leaves in the concept semantic network have reverse topological level 1. After removing all these leaves, all new leaves in the new network have reverse topological order 2. Iteratively applying this approach, we can determine a reverse topological level for all concept semantic types. When one concept has multiple semantic types, we assign the concept a semantic type closest to the leaves of the concept semantic network. Under this measurement, two concepts are likely to be close if there are many short and concrete paths between them.

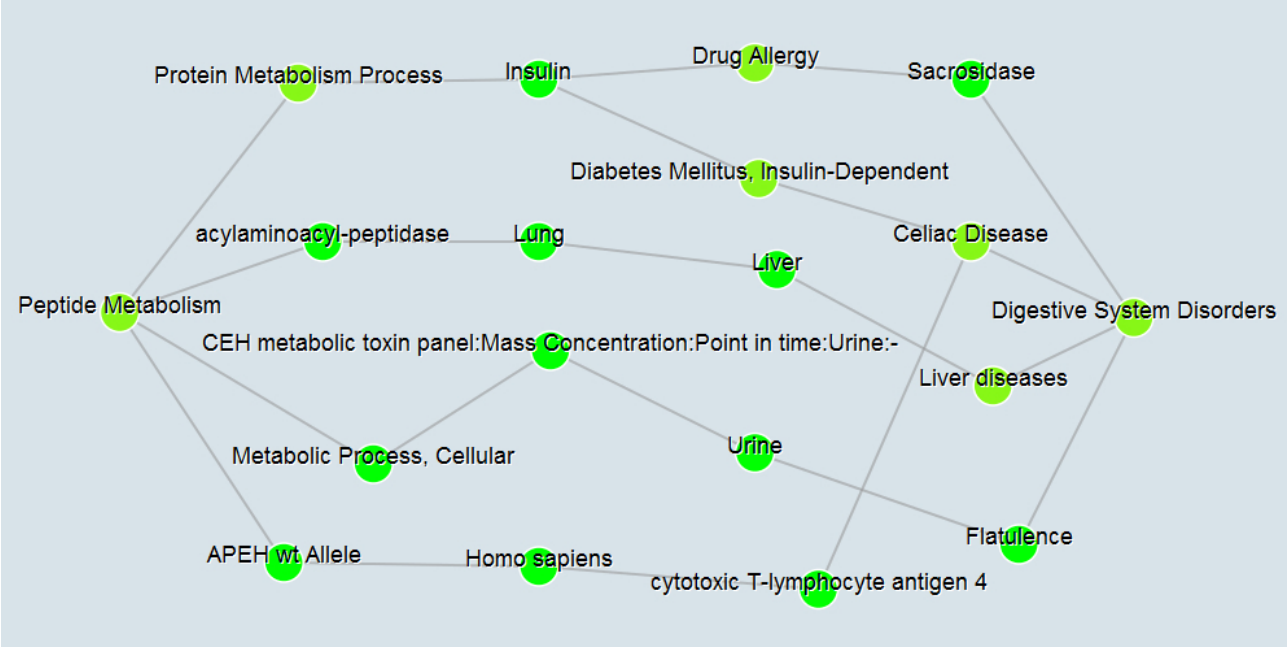
Results

Transitive Relationship Queries and Visualization

onGrid supports both graph queries and conceptual relationship studies on UMLS data sources. For graph queries, it supports distance and shortest path queries on a conceptual network built upon UMLS data sources. To use this function, users can input a start biomedical concept (or CUI) and an end biomedical concept (or CUI), in which the system will output shortest paths visualized in a network structure. Figure 2 provides an illustration of such a network of structured paths between Peptide Metabolism (Semantic Type: Molecular Function) and Digestive System Disorders (Semantic Type: Disease or Syndrome). Users can choose to see an edge's semantic type by moving their mouse to an edge (eg, Sacrosidase—"may_treat"—Digestive System Disorders), or simply selecting the option to show all of them.

In the current version, the basic settings of onGrid, including neighborhood search range, sink and source vertex handling, and semantic restrictions, follow our preliminary study [16], which demonstrates that this setting is cost-effective for knowledge discovery on the UMLS. In this setting, since k is configured to be 6, the system guarantees finding exact distances no more than 6 hops, or at least one shortest path no more than 6 hops, on the underlying graph built upon the selected UMLS data sources.

Figure 2. An illustration of visualized paths between two biomedical concepts.



Large Scale Relationship Matrix Generation

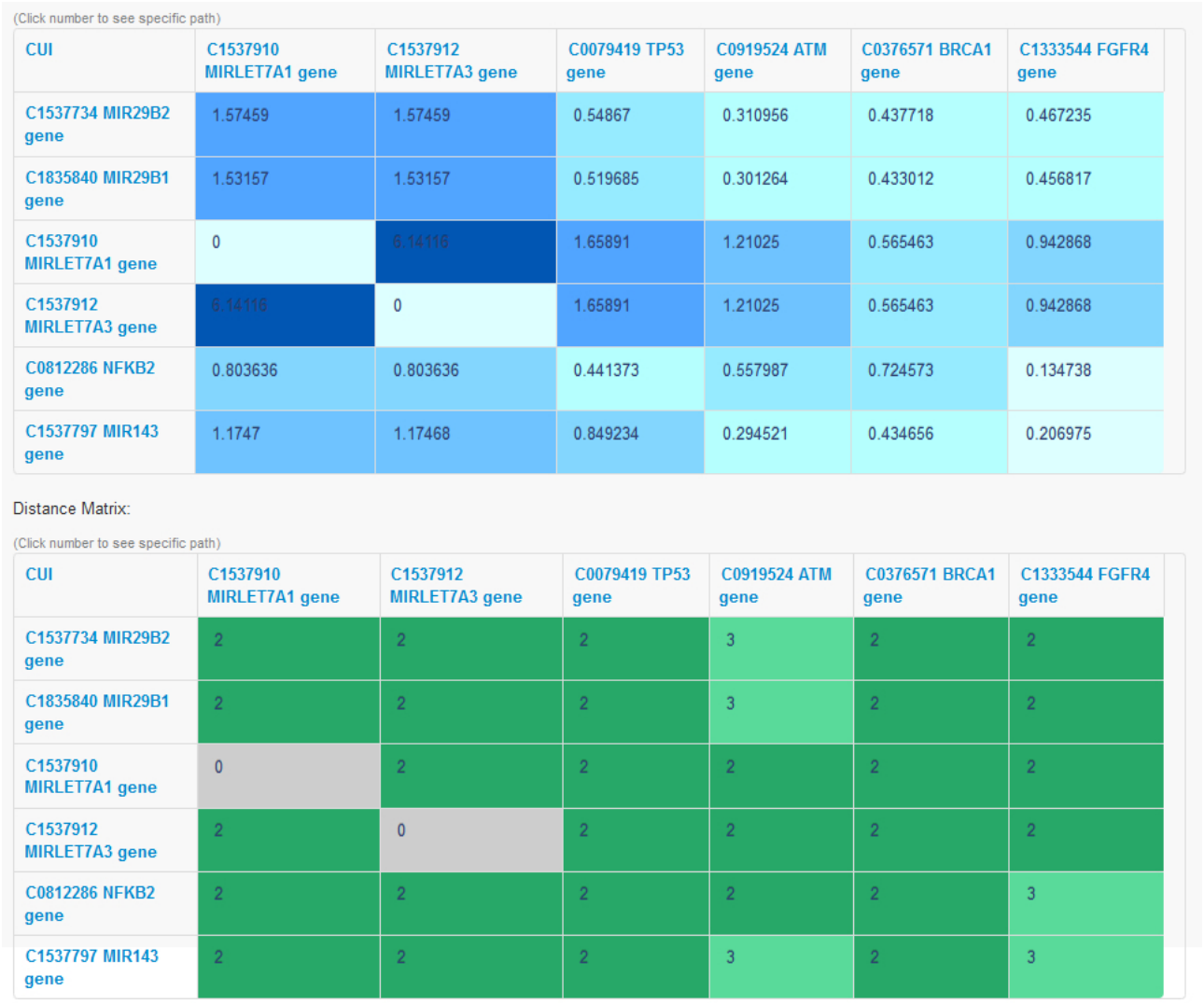
In addition to the path queries, onGrid supports powerful conceptual relationship studies by allowing users to input two sets of biomedical concepts (or CUIs) and builds a distance heatmap/matrix as well as a relationship heatmap/matrix (as illustrated in Figure 3).

The distance heatmap provides a distance between every two concepts. However, distance alone may not be a good measurement for the relationships between medical concepts. Thus, onGrid provides the relationship heatmap using the concept relationship measurement function $R(u,v)$ defined above, which extends the measurement in [16] by giving more weight to concrete paths, that is, paths with fewer abstract concepts. Similar to [16], paths with only one edge (ie, direct relations) are not counted in $R(u,v)$ to avoid bias towards existing knowledge. Below we examine a large scale study on conceptual relationships between medical terms that uses the relationship matrix generated under this measurement. Finally, onGrid provides a very convenient feature for exploring these two matrices: If users are interested in any particular pair of CUIs, they can click the corresponding unit and onGrid will provide the result for the shortest path query on those two medical concepts.

onGrid also supports studies on large sets of medical concepts for users who wish to use this functionality due to the large amount of processing time required. onGrid supports these types of applications by allowing users to upload files, track their

jobs, and download the results (a valid email address is required for these purposes).

Figure 3. An example of relationship matrix and distance matrix generated by onGrid (in the relationship matrix, a higher number means a closer relationship).



(Click number to see specific path)

CUI	C1537910 MIRLET7A1 gene	C1537912 MIRLET7A3 gene	C0079419 TP53 gene	C0919524 ATM gene	C0376571 BRCA1 gene	C1333544 FGFR4 gene
C1537734 MIR29B2 gene	1.57459	1.57459	0.54867	0.310956	0.437718	0.467235
C1835840 MIR29B1 gene	1.53157	1.53157	0.519685	0.301264	0.433012	0.456817
C1537910 MIRLET7A1 gene	0	6.14116	1.65891	1.21025	0.565463	0.942868
C1537912 MIRLET7A3 gene	6.14116	0	1.65891	1.21025	0.565463	0.942868
C0812286 NFKB2 gene	0.803636	0.803636	0.441373	0.557987	0.724573	0.134738
C1537797 MIR143 gene	1.1747	1.17468	0.849234	0.294521	0.434656	0.206975

Distance Matrix:

(Click number to see specific path)

CUI	C1537910 MIRLET7A1 gene	C1537912 MIRLET7A3 gene	C0079419 TP53 gene	C0919524 ATM gene	C0376571 BRCA1 gene	C1333544 FGFR4 gene
C1537734 MIR29B2 gene	2	2	2	3	2	2
C1835840 MIR29B1 gene	2	2	2	3	2	2
C1537910 MIRLET7A1 gene	0	2	2	2	2	2
C1537912 MIRLET7A3 gene	2	0	2	2	2	2
C0812286 NFKB2 gene	2	2	2	2	2	3
C1537797 MIR143 gene	2	2	2	3	2	3

Validating onGrid by Studying Conceptual Relationships Between Diseases

onGrid can be applied to study the relations in a set of medical concepts or between two sets of medical concepts. To carry out the study, one can map these concepts to ontology terms in the UMLS using the natural language processing method described above and then generate a relationship matrix for these terms. In order to crossvalidate our results by an available source, we used onGrid to study the disease relationships in the Online Mendelian Inheritance in Man (OMIM) ontology dataset, which is a database collection of diseases with a genetic component. First, we use onGrid (on the full UMLS data source configuration) to generate a relationship matrix between diseases in OMIM and genes in the Human Genome Organization (HUGO). Then, given a threshold δ, we are able to convert the relationship matrix into a 0-1 relationship matrix. We construct weighted relations T over OMIM diseases by the number of genes shared by two diseases in the 0-1 relationship matrix. To crossvalidate our results, we build the same weighted disease relations S on the Comparative Toxicogenomics Database (CTD) [20]. We use fold enrichment to measure our results. The fold enrichment function is defined as $f(\alpha) = (|S'(\alpha)|/|S'|)/(|T(\alpha)|/|T|)$ where $S' = S \cap T$; $S'(\alpha)$ is the number of elements in S that are ranked in the top α percent of T according to the weight of disease pairs; $T(\alpha)$ is the number of elements in T that are ranked in the top α percent of T. It is quite intuitive that $f(\alpha)$ will be close to 1 if T is random, and if $f(\alpha)$ is much larger than 1, it suggests that T is statistically significant with respect to S.

Here we give a small hypothetical example to illustrate the above fold enrichment measurement. Let T = (<A,B> , <A,C> , <B,D> , <E,F> , <A,E> , <B,C> , <B,E> , <D,E> , <D,F> , <C,D>), which contains 10 pairs of diseases ranked in the descending order of their closeness. Let S = {<A,C>, <B,D>, <A,E>, <E,F>, <H,G>, <E,H>}, which contains six pairs of confirmed disease pairs. Then $S' = S \cap T$ = {<A,C>, <B,D>, <A,E>, <E,F>}, and $S'(\alpha = 20\%)$ = {<A,C>}. Thus, the fold enrichment at $\alpha = 20\%$ is $f(\alpha = 20\%) = (|S'(\alpha)|/|S'|)/(|T(\alpha)|/|T|) = (1/4)/(2/10) = 1.25$, and the maximum fold enrichment $f(\alpha) = 3.75$ (when $\alpha = 40\%$).

The fold enrichment results of the OMIM disease relationships generated by onGrid with respect to CTD are provided in Figures 4 and 5. To understand the advantage of onGrid over kDLS, we also include the kDLS in the study.

From Figures 4 and 5, we can see that fold enrichment values are much larger than 1. They generally increase when the threshold δ increases. This is because when the threshold δ is high, only the disease pairs sharing the most genes (ie, most significant disease-disease pairs) are left for study. Thus, to avoid studying too few disease pairs, the thresholds in this study were set to an upper limit. We also noticed that these values get smaller when percentage α increases. This is understandable because according to the definition, when α increases, the difference between the numerator and denominator tends to get smaller, and $f(\alpha) = 1$ when $\alpha = 100$. These fold enrichment tests suggest that the disease pair results obtained by onGrid are statistically significant in the crossvalidation with an external dataset, CTD.

In addition, Figures 4 and 5 include corresponding results generated from the original kDLS algorithm (indicated by dashed lines). To ensure the results are comparable, the percentiles of relationships (ie, entries) for δ thresholds in the onGrid matrices were obtained and used to determine appropriate δ values for the kDLS matrices. Table 1 lists their respective δ values for each threshold level. onGrid tends to generate higher fold enrichment values for each respective α, suggesting that incorporating semantic types leads to more focused and correlated diseases and genes.

Table 1. Corresponding thresholds δ for kDLS and onGrid.

Threshold level	δ for onGrid	δ for kDLS
1	0.45	0.73
2	0.5	0.8
3	0.55	0.87
4	0.6	0.93
5	0.65	1
6	0.7	1.08
7	0.75	1.15
8	0.8	1.23
9	0.85	1.29
10	0.9	1.35
11	0.95	1.41
12	1	1.48
13	1.05	1.53
14	1.1	1.59
15	1.15	1.68
16	1.2	1.74
17	1.25	1.82
18	1.3	1.9
19	1.35	2
20	1.4	2.07

Figure 4. Maximum Fold Enrichment for both kDLS and onGrid.

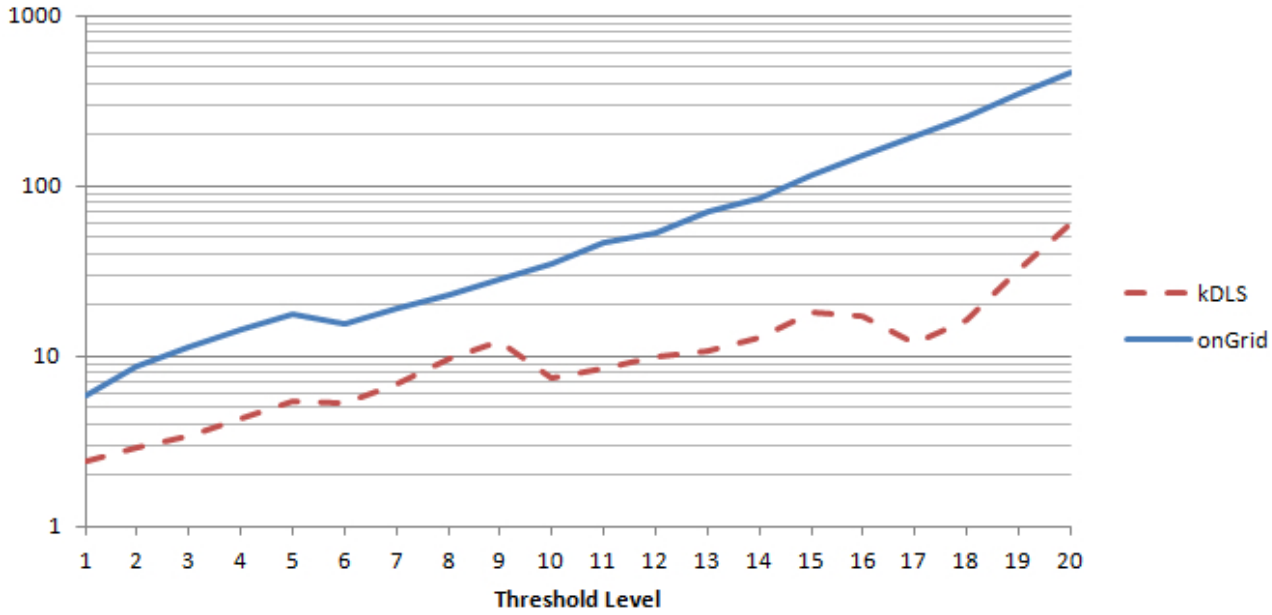
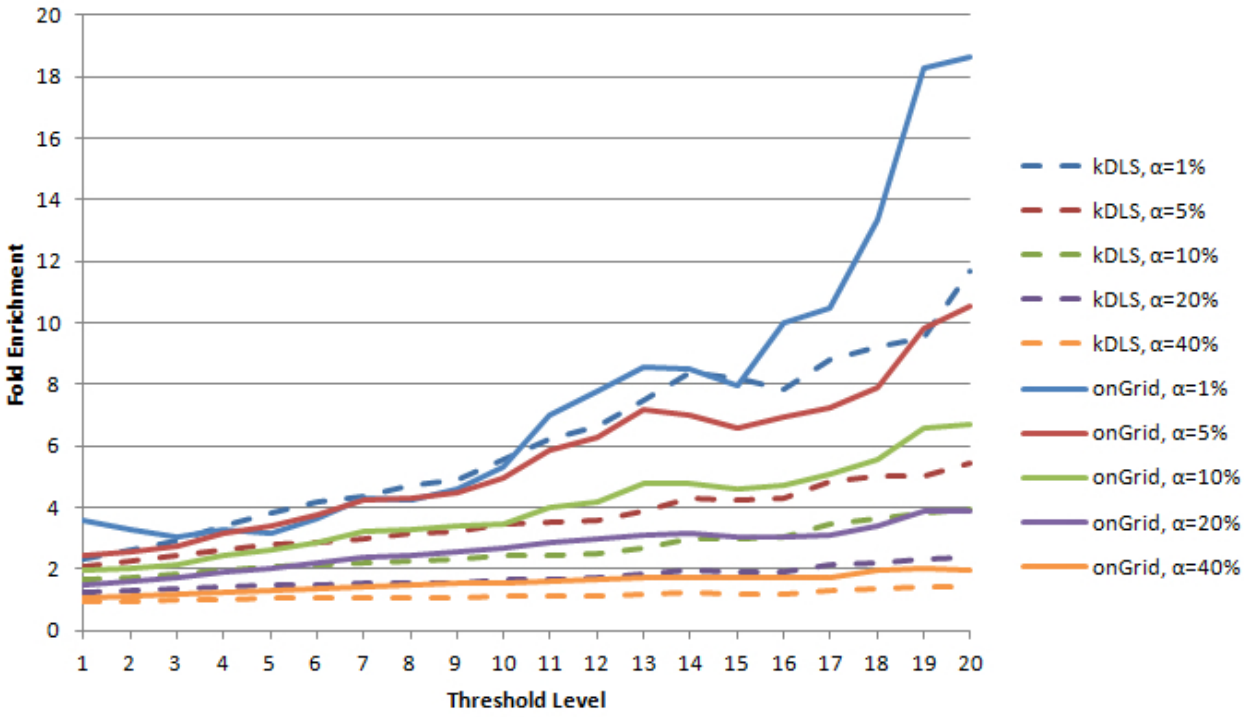


Figure 5. Fold Enrichment f(α) for both kDLS and onGrid.



Comparing onGrid and Comparative Toxicogenomics Database on Conceptual Disease Relationships

We are able to further study any interested diseases to observe other diseases most related to them. For demonstration purposes, we use adenocarcinoma of lung and glioblastoma in this study. The relationship between two diseases is measured by the number of genes shared between them. This measurement can be used to study the disease relationships in both onGrid results and in the CTD. According to the role of the threshold δ, one can infer that when δ decreases the differences among relationships (ie, edge thickness) blur, and when δ increases the differences among relationships become sharp, and at some point only the thickest edges will show. For conciseness in this paper, we show only the results under δ = 1.1 as a balanced result of the two effects. The top-ranked diseases related to the two diseases are presented in Figures 6 and 7 in circular arc graphs. The edges (relationships) connected to the studied diseases (adenocarcinoma of lung or glioblastoma) are shown in red, and other edges are shown in gray. An edge thickness is proportional to the normalized edge weight, which is obtained by categorizing the number of shared genes into 10 levels.

To demonstrate the advantage of onGrid, we also conducted the same analysis using CTD (Figures 8 and 9).

In Figures 6-9, we can see that the disease relationships generated by onGrid have a larger weight variation (visualized by the thickness of edges) compared to the disease relationships of CTD. Thus, it is easier to distinguish closeness between diseases in onGrid than CTD. In addition, the top-related diseases by onGrid (Figures 6 and 7) are mostly leukemia and carcinoma for adenocarcinoma of lung, and mostly carcinoma for glioblastoma. They are consistent with the disease mechanisms contained in the UMLS ontologies. Furthermore, we found that other independent studies partially confirm the results generated by onGrid. For example, the loss of heterozygosity on chromosome 3p was observed for both patients of small cell carcinoma of lung and patients of adenocarcinoma of lung [21], validating their relationships revealed by onGrid. As another example, lymphoma, a top-related disease to adenocarcinoma of lung by onGrid, was observed to have the same effect with adenocarcinoma of lung in the combined inactivation of oncogenes MYC and K-ras in a study using mouse models [22]. Similarly, we also found studies between glioblastoma and top diseases related to glioblastoma by onGrid. In contrast, the top-related diseases by CTD (Figures 8 and 9) are quite general, mostly reflecting the toxicology viewpoints of liver necrosis and kidney damage. These observations suggest that onGrid provides important information for studying the conceptual relationships between diseases.

Figure 6. Top diseases related to adenocarcinoma of lung by onGrid ($\delta = 1.1$).

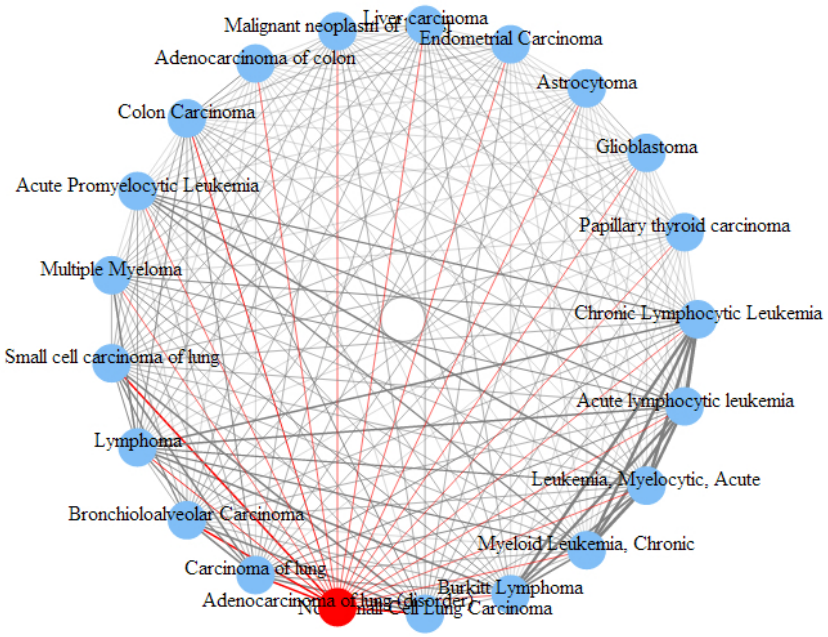


Figure 7. Top diseases related to glioblastoma by onGrid (δ = 1.1).

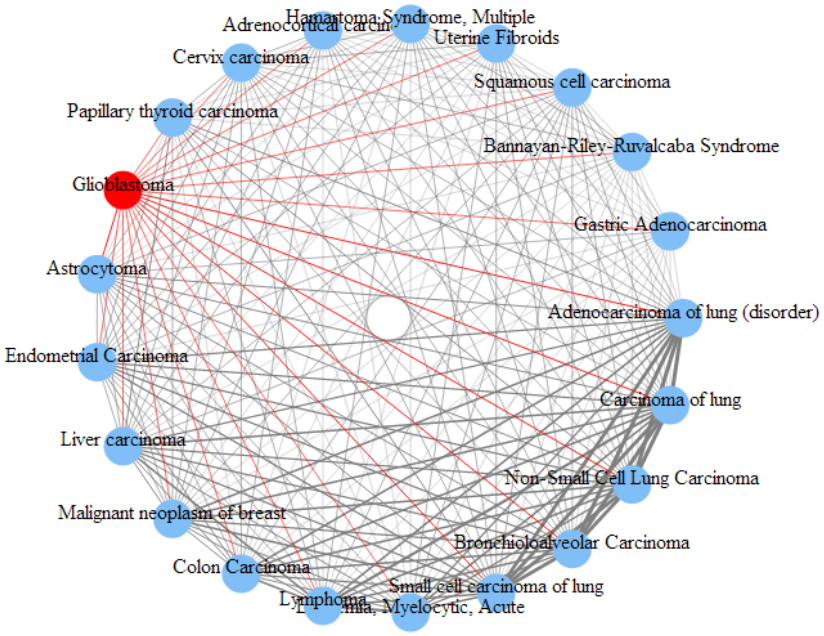


Figure 8. Top diseases related to adenocarcinoma of lung according to CTD.

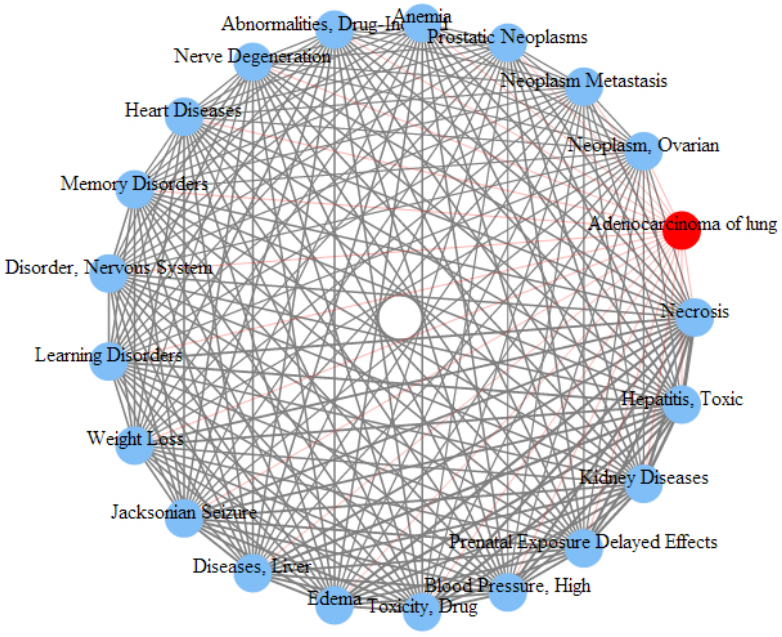
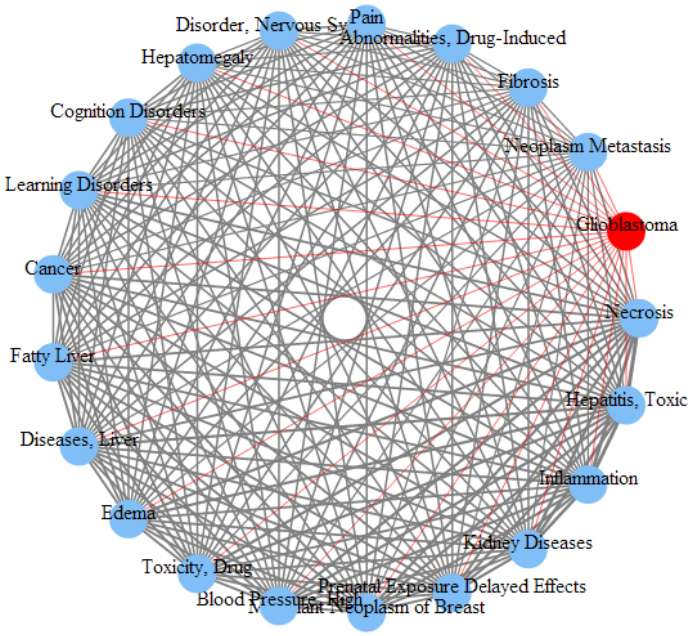


Figure 9. Top diseases related to glioblastoma according to CTD.



Discussion

Studying Concept Relationships Using onGrid

Above we have shown the effectiveness of using onGrid for studying disease-disease relationships. These results can be used to assist other studies such as analyzing electronic health records. In addition, onGrid can be used for studying many conceptual relationships other than disease-disease or disease-gene relationships. We can use onGrid to study the relationships among many important biomedical concepts, including drugs, diseases, genes, side effects, etc. To perform these studies, we may use corresponding ontologies such as RxNorm (for drugs), International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) (for diseases), OMIM (for diseases with a genetic component), Gene Ontology (GO) (for genes), and Medical Dictionary for Regulatory Activities (MedDRA) (for side effects). These studies can be used to assist many biomedical applications, such as identifying drug side effects and drug repurposing candidates. We can further leverage these studies with research on external datasets or ontologies.

Limitations of the Conceptual Relationship Study Using Unified Medical Language System

Since UMLS is a collection of ontologies, it is essentially a body of knowledge. Although knowledge discovery on such data will produce transitive associations that may not have been noticed before, it will not produce knowledge that is out of the given ontological data. Consequently, the discovered relationships are likely to concentrate on well-studied concepts. In addition, since UMLS does not provide a weight on the concept relationships, it is not clear how important a relationship is. Thus, a transitive relationship on the UMLS may not be reliable. onGrid provides an advanced heuristic solution by considering both the discovered paths and semantic types. The crossvalidation demonstrates that the discovered results are statistically significant in aggregation. However, for one individual relationship between two concepts, it is difficult to further identify its statistical significance with the given resource in the UMLS. To complement this disadvantage, onGrid provides the path query function for two concepts and visualizes the discovered paths. Thus, domain experts are able to manually verify the validity of the transitive relationships between them. We expect that, in the future, by integrating information from external data sources, we will be able to perform efficient conceptual relationship studies that exceed the limitation of UMLS.

Conclusions

onGrid provides an efficient Web-based platform to perform conceptual relationship studies using the UMLS. The current version of onGrid uses graph indexing with semantic relations as its server side index engine and can be easily upgraded in the future. onGrid can efficiently output shortest paths between two medical concepts as well as build relationship and distance heatmaps. The relationship heatmap enables researchers to quickly identify highly related medical concepts and directly check the transitive relation between any two concepts on the heatmap by clicking the corresponding unit. Our study on the conceptual relationships between OMIM diseases demonstrates the effectiveness of using onGrid in studying medical concept relations. We expect onGrid will be used for many applications to assist conceptual relationship studies in the biomedical field.

Conflicts of Interest

None declared.

References

1. Swanson D. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 1986;30(1):7-18. [Medline: 3797213]
2. Hristovski D, Peterlin B, Mitchell J, Humphrey S. Using literature-based discovery to identify disease candidate genes. Int J Med Inform 2005 Mar;74(2-4):289-298. [doi: 10.1016/j.ijmedinf.2004.04.024] [Medline: 15694635]
3. Ahlers C, Hristovski D, Kilicoglu H, Rindflesch T. Using the literature-based discovery paradigm to investigate drug mechanisms. AMIA Annu Symp Proc 2007:6-10 [FREE Full text] [Medline: 18693787]
4. Hristovski D, Rindflesch T, Peterlin B. Using literature-based discovery to identify novel therapeutic approaches. Cardiovasc Hematol Agents Med Chem 2013 Mar;11(1):14-24. [Medline: 22845900]
5. Petric I, Urbancic T, Cestnik B, Macedoni-Luksic M. Literature mining method RaJoLink for uncovering relations between biomedical concepts. J Biomed Inform 2009 Apr;42(2):219-227 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.004] [Medline: 18771753]
6. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
7. Hristovski D, Friedman C, Rindflesch T, Peterlin B. Exploiting semantic relations for literature-based discovery. AMIA Annu Symp Proc 2006:349-353 [FREE Full text] [Medline: 17238361]
8. Wilkowski B, Fiszman M, Miller C, Hristovski D, Arabandi S, Rosemblat G, et al. Graph-based methods for discovery browsing with semantic predications. AMIA Annu Symp Proc 2011;2011:1514-1523 [FREE Full text] [Medline: 22195216]
9. Lee W, Shah N, Sundlass K, Musen M. Comparison of ontology-based semantic-similarity measures. AMIA Annu Symp Proc 2008:384-388 [FREE Full text] [Medline: 18999312]
10. Melton G, Parsons S, Morrison F, Rothschild A, Markatou M, Hripcsak G. Inter-patient distance metrics using SNOMED CT defining relationships. J Biomed Inform 2006 Dec;39(6):697-705 [FREE Full text] [doi: 10.1016/j.jbi.2006.01.004] [Medline: 16554186]
11. Payne P, Payne P, Borlawsky T, Borlawsky T, Kwok A, Greaves A, et al. Supporting the design of translational clinical studies through the generation and verification of conceptual knowledge-anchored hypotheses. AMIA Annu Symp Proc 2008:566-570 [FREE Full text] [Medline: 18998958]
12. Payne P, Kwok A, Dhaval R, Borlawsky T. Conceptual dissonance: evaluating the efficacy of natural language processing techniques for validating translational knowledge constructs. Summit on Translat Bioinforma 2009;2009:95-99 [FREE Full text] [Medline: 21347178]
13. McInnes B, Pedersen T, Pakhomov S. UMLS-Interface and UMLS-Similarity : open source software for measuring paths and semantic similarity. AMIA Annu Symp Proc 2009;2009:431-435 [FREE Full text] [Medline: 20351894]
14. Nguyen HA, Al-Mubaid H. New ontology-based semantic similarity measure for the biomedical domain. 2006 Presented at: IEEE International Conference on Granular Computing; May 10-12, 2006; Atlanta, GA. [doi: 10.1109/GRC.2006.1635880]
15. Xiang Y, Payne P, Huang K. Transactional database transformation and its application in prioritizing human disease genes. IEEE/ACM Trans Comput Biol Bioinform 2012;9(1):294-304 [FREE Full text] [doi: 10.1109/TCBB.2011.58] [Medline: 21422495]
16. Xiang Y, Lu K, James S, Borlawsky T, Huang K, Payne P. k-Neighborhood decentralization: a comprehensive solution to index the UMLS for large scale knowledge discovery. J Biomed Inform 2012 Apr;45(2):323-336 [FREE Full text] [doi: 10.1016/j.jbi.2011.11.012] [Medline: 22154838]
17. Jin R, Ruan N, Xiang Y, Lee V. A Highway-Centric Labeling Approach for Answering Distance Queries on Large Sparse Graphs. 2012 Presented at: ACM SIGMOD International Conference on Management of Data; May 20-24, 2012; Scottsdale, AZ p. 445-456. [doi: 10.1145/2213836.2213887]
18. Raghavan P, Fosler-Lussier E, Lai AM. Exploring semi-supervised coreference resolution of medical concepts using semantic and temporal features. 2012 Presented at: Conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies; 2012; Montréal, QC p. 731-741.
19. Ren K, Lai A, Mukhopadhyay A, Machiraju R, Huang K, Xiang Y. Effectively processing medical term queries on the UMLS Metathesaurus by layered dynamic programming. BMC Med Genomics 2014;7 Suppl 1:S11 [FREE Full text] [doi: 10.1186/1755-8794-7-S1-S11] [Medline: 25079259]
20. Davis A, King B, Mockus S, Murphy C, Saraceni-Richards C, Rosenstein M, et al. The Comparative Toxicogenomics Database: update 2011. Nucleic Acids Res 2011 Jan;39(Database issue):D1067-D1072 [FREE Full text] [doi: 10.1093/nar/gkq813] [Medline: 20864448]
21. Yokota J, Wada M, Shimosato Y, Terada M, Sugimura T. Loss of heterozygosity on chromosomes 3, 13, and 17 in small-cell carcinoma and on chromosome 3 in adenocarcinoma of the lung. Proc Natl Acad Sci U S A 1987 Dec;84(24):9252-9256 [FREE Full text] [Medline: 2892196]
22. Tran P, Fan A, Bendapudi P, Koh S, Komatsubara K, Chen J, et al. Combined Inactivation of MYC and K-Ras oncogenes reverses tumorigenesis in lung adenocarcinomas and lymphomas. PLoS One 2008;3(5):e2125 [FREE Full text] [doi: 10.1371/journal.pone.0002125] [Medline: 18461184]

Abbreviations

BFS: breadth-first search
BITOLA: biomedical discovery support system
CTD: Comparative Toxicogenomics Database
CUI: concept unique identifier
DFS: depth-first search
GO: Gene Ontology
HUGO: Human Genome Organization
ICD-9-CM: International Classification of Diseases, 9th Revision, Clinical Modification
kDLS: K-neighborhood Decentralization Labeling Scheme
LDPMap: Layered Dynamic Programming Map
MedDRA: Medical Dictionary for Regulatory Activities
OMIM: Online Mendelian Inheritance in Man
onGrid: Online conceptual study platform using Graph indexing
UMLS: Unified Medical Language System
UTS: UMLS Terminology Services

Edited by G Eysenbach; submitted 11.03.14; peer-reviewed by I Yoo, S Lim Choi Keung, A Zahrawi; comments to author 04.08.14; revised version received 15.08.14; accepted 16.08.14; published 07.10.14.

Please cite as:
Albin A, Ji X, Borlawsky TB, Ye Z, Lin S, Payne PRO, Huang K, Xiang Y
Enabling Online Studies of Conceptual Relationships Between Medical Terms: Developing an Efficient Web Platform
JMIR Med Inform 2014;2(2):e23
URL: http://medinform.jmir.org/2014/2/e23/
doi:10.2196/medinform.3387
PMID:25600290

Original Paper

Return on Investment in Electronic Health Records in Primary Care Practices: A Mixed-Methods Study

Yeona Jang[1*], MSc, MBA, PhD; Michel A Lortie[2*], Ing; Steven Sanche[2], MSc

[1]McGill University, Desautels Faculty of Management, Montreal, QC, Canada

[2]St Mary's Research Centre, Montreal, QC, Canada

[*]these authors contributed equally

Corresponding Author:
Yeona Jang, MSc, MBA, PhD
McGill University
Desautels Faculty of Management
1001 Rue Sherbrooke Ouest
Montreal, QC, H3A 1G5
Canada
Phone: 1 514 398 8489
Fax: 1 514 398 3876
Email: yeona.jang@mcgill.ca

Abstract

Background: The use of electronic health records (EHR) in clinical settings is considered pivotal to a patient-centered health care delivery system. However, uncertainty in cost recovery from EHR investments remains a significant concern in primary care practices.

Objective: Guided by the question of "When implemented in primary care practices, what will be the return on investment (ROI) from an EHR implementation?", the objectives of this study are two-fold: (1) to assess ROI from EHR in primary care practices and (2) to identify principal factors affecting the realization of positive ROI from EHR. We used a break-even point, that is, the time required to achieve cost recovery from an EHR investment, as an ROI indicator of an EHR investment.

Methods: Given the complexity exhibited by most EHR implementation projects, this study adopted a retrospective mixed-method research approach, particularly a multiphase study design approach. For this study, data were collected from community-based primary care clinics using EHR systems.

Results: We collected data from 17 primary care clinics using EHR systems. Our data show that the sampled primary care clinics recovered their EHR investments within an average period of 10 months (95% CI 6.2-17.4 months), seeing more patients with an average increase of 27% in the active-patients-to-clinician-FTE (full time equivalent) ratio and an average increase of 10% in the active-patients-to-clinical-support-staff-FTE ratio after an EHR implementation. Our analysis suggests, with a 95% confidence level, that the increase in the number of active patients (P=.006), the increase in the active-patients-to-clinician-FTE ratio (P<.001), and the increase in the clinic net revenue (P<.001) are positively associated with the EHR implementation, likely contributing substantially to an average break-even point of 10 months.

Conclusions: We found that primary care clinics can realize a positive ROI with EHR. Our analysis of the variances in the time required to achieve cost recovery from EHR investments suggests that a positive ROI does not appear automatically upon implementing an EHR and that a clinic's ability to leverage EHR for process changes seems to play a role. Policies that provide support to help primary care practices successfully make EHR-enabled changes, such as support of clinic workflow optimization with an EHR system, could facilitate the realization of positive ROI from EHR in primary care practices.

(JMIR Med Inform 2014;2(2):e25) doi:10.2196/medinform.3631

KEYWORDS

return on investment in electronic health records; cost recovery from EHR implementation; ROI indicator; physician satisfaction with EHR; primary care practices

Introduction

Context

The use of electronic health records (EHR) in clinical settings is widely recommended as an innovation enabler with potential benefits of reducing health care costs, while improving quality and safety, and is considered central to achieving patient-centered health care [1-4]. As a wide array of EHR projects have been implemented within various health care settings, the health care field is rich with volumes of work examining the benefits of EHR. However, the existing literature reports mixed results in benefits realized from EHR implementation [5,6]. Such mixed results suggest that the implementation of EHR systems does not automatically guarantee the conversion of potential benefits into realized benefits.

The implementation of EHR systems within primary care practices is seen as particularly complex [7-10], with physicians and other staff in primary care practices citing obstacles such as difficulty in adapting to the significant changes in workflow and the time commitment required to learn to use the new software while prioritizing patient care [11-14]. While there is a growing body of evidence that EHR can be a valuable tool for improving quality of care and patient safety with relatively positive perceptions about EHR benefits [15-17], uncertainty about cost recovery of an EHR investment remains a significant concern in primary care practices [7,8,18,19]. Various studies on EHR impact and adoption also raise the need for cost-benefit analysis of EHR investments [5,20]. Thus, this study seeks to assess the return on investment (ROI) from an EHR implementation in primary care settings, aiming to complement the current insights on cost recovery concerns in existing literature.

Measurement

Return on investment is a common approach to measuring rates of return on money invested, in terms of increased profit attributable to the investment. A standard ROI is defined as follows:

ROI = (Gain from investment - Cost of investment)/Cost of investment

Results reported by various studies regarding ROI from EHR systems in primary care settings are mixed [21-25]. Most of the existing literature used a bottom-up approach identifying specific cost-saving areas and collecting the data on financial savings made in these areas attributable to EHR systems. However, EHR is a process-enabling information technology (IT) that offers the opportunity to streamline information-intensive workflow, remove manual hand-off of data and information, and facilitate coordination—thus facilitating the execution of entire business processes rather than individual tasks. Due mainly to the context-sensitive nature of benefits realization from a process-enabling IT such as EHR and the scarcity of detailed financial data relating to gains and/or savings directly attributable to an EHR system in primary care clinics, this study used break-even-point analysis as an indicator of ROI, instead of standard ROI analysis.

The break-even point of an EHR investment is defined as the number of months it takes a clinic to recover the cost of the EHR system and other associated implementation costs, with increased revenues and/or decreased expenses. Increases in revenues and/or decreases in expenses are assessed by considering net revenues during three distinct periods of time: pre-EHR, peri-EHR, and post-EHR. The pre-EHR period is defined as the full fiscal year before the implementation of an EHR system started. The peri-EHR period is defined as the fiscal year(s) containing the EHR implementation period (ie, during EHR implementation). If the peri-EHR period covers more than one fiscal year, the net revenue is averaged over these fiscal years. The post-EHR period is defined as the full fiscal year following the end of the peri-EHR period.

To calculate the break-even point of implementing an EHR system in a clinic, the cost of EHR implementation is set equal to the difference in the clinic's net revenue between the pre-EHR and peri-EHR periods, plus the difference in the clinic's net revenue between the pre-EHR and post-EHR periods, as summarized in the following formula:

$$C_{EHR}=[(NR_{peri}-NR_{pre})/12]*M_{imp} + [(NR_{post}-NR_{pre})/12] * (M_{break\text{-}even}-M_{imp})$$

In this formula: C_{EHR}=cost of EHR implementation, NR_{peri}=annual clinic net revenue in the peri-EHR period, NR_{pre}=annual clinic net revenue in the pre-EHR period, NR_{post}=annual clinic net revenue in the post-EHR period, M_{imp}=the number of months taken to complete the EHR implementation in the clinic, and $M_{break\text{-}even}$=months to break even. The net revenue of a clinic is defined as the sum of the physicians' billings for work done in the clinic minus any expenses that the clinic pays to maintain its ongoing practice. In the case where the months to break even were less than the months of EHR implementation, in other words, the net revenue difference between pre-EHR and peri-EHR periods is large enough to recover the cost of EHR implementation, the formula was adjusted by setting the cost of EHR implementation equal to the difference in a clinic's net revenue between the pre-EHR and peri-EHR periods, or:

$$C_{EHR} =[(NR_{peri} - NR_{pre})/12]* M_{break\text{-}even}$$

Objectives

Guided by the research question "When implemented in primary care practices, what will be the ROI from an EHR implementation?", the objectives of this research are twofold: (1) to assess the ROI from an EHR implementation in primary care practices by measuring the time required to recover the cost of converting a clinic from a paper-based environment to an EHR-enabled environment and (2) to identify principal factors affecting the realization of a positive ROI from an EHR implementation in primary care practices. Such ROI information related to cost recovery of an EHR investment would be helpful to both clinics considering implementing EHR systems and to policy makers designing EHR-adoption funding programs and policies.

Methods

Sample

Community-based, primary care clinics meeting the following four eligibility criteria were recruited for this study on ROI from EHR in primary-care settings. First, this study focused on community-based, primary care clinics. Thus, specialty care clinics and walk-in clinics were excluded. Second, clinics were required to have implemented EHR systems. Third, clinics were required to have been paper-based in the past, in order to ensure that the comparison between pre-EHR and post-EHR implementation performance was possible for the ROI calculation. Fourth, clinics were required to provide operational and financial data necessary to calculate ROI, as well as the information on challenges and opportunities that they had experienced both during and after the EHR implementation.

The research team contacted 132 randomly selected community-based, primary care clinics in Canada that met the first two eligibility criteria for recruitment to the study. Of the 132 clinics, 62 clinics declined to participate, mostly citing time constraints. Of the 70 clinics remaining, 34 clinics were not eligible, mainly because they were unable to provide the operational and financial data necessary to calculate ROI. Of the 36 eligible clinics, 19 clinics later declined to participate, due mainly to time constraints. Thus, data were collected from a total of the 17 eligible clinics, resulting in the study participation rate of 13%, which is relatively consistent with typical participation rates of family physicians reported in other studies involving interviews and observations [26]. No statistically significant differences were observed between participating and non-participating primary care clinics in terms of geographic location (P=.315), the number of physicians or other clinicians (P<.001), or the number of patients per physician (P=.192). Table 1 summarizes the basic statistics on the size of the sampled clinics. We used Full Time Equivalent (FTE) in comparing the size of primary care clinics.

Table 1. Basic statistics on the size of a primary care clinic in the study.

	Average	SD	Median	Minimum	Maximum
Clinic size: clinician FTEs					
Pre-EHR period	3.4	2.6	3.0	1.0	8.5
Post-EHR period	3.6	2.4	3.0	0.8	8.0
Clinic size: clinical support staff FTEs					
Pre-EHR period	3.4	2.9	2.8	1.0	12.0
Post-EHR period	4.2	3.1	3.0	0.9	12.0

Methodology

Given the complexity exhibited by most EHR implementation projects, this study used a mixed-method research approach, particularly a multiphase study design [27]. By combining quantitative and qualitative data, mixed-method research can provide a fuller understanding of the complex and multidimensional world of primary care clinics than would otherwise be achieved using either approach alone.

In the quantitative study phase, questionnaire modules were designed, based on prior research in the existing literature [28-33], to collect data on EHR implementation costs, EHR functionalities in use, physician satisfaction with EHR, and physicians' perceptions about the impact of EHR on operational efficiency and on quality of care. Each clinic respondent was required to complete the study instruments using the online questionnaires or researcher-assisted telephone questionnaires. The minimum financial data required for the study include clinic revenue and clinic net revenue as well as EHR implementation cost that consisted of EHR software costs, hardware costs, support costs, and labor costs associated with EHR system implementation and training, in the three different periods—before EHR implementation, during EHR implementation, and after EHR implementation. The minimum operational data required for this ROI study include the number of active patients, clinician FTEs, and clinical support staff FTEs, in the same three periods. The lead researcher served as dedicated support liaison for clinics, in order to ensure that the costs of the EHR implementation, as well as other financial and operational data before and after EHR implementation, were abstracted from clinic records in a consistent fashion. In the subsequent qualitative study, semistructured interviews and observations were conducted with clinic staff and physicians identified as responsible for such functions as patient appointment management, patient record management, test results management, patient encounters, and billing, to assess factors affecting the realization of a positive ROI from an EHR implementation in primary care practices.

The data collected from 17 sampled primary care clinics were documented and analyzed using statistical analysis and grounded theory [34]. As break-even points were analyzed, we compared those clinics that were successful in realizing a positive ROI from EHR implementation to those that were less successful, in an attempt to identify principal factors impacting the realization of positive ROI from EHR. In particular, we used linear regression analysis to estimate the relationships of the outcome variable "break-even point" with the explanatory variables that include the codes identified from the qualitative data through the coding process.

Results

Analysis of Break-Even Point as Indicator of Return on Investment

Overview

Our analysis suggests that the sampled primary care clinics typically recovered their investment in EHR within an average of 10 months (95% confidence interval: 6.2 months, 17.4 months), seeing more patients with improved active-patients-to-clinician-FTE and active-patients-to-clinical-support-staff-FTE ratios in the post-EHR implementation period.

Change in Clinic Net Revenue After Implementation of Electronic Health Records

Once an EHR system is implemented, a key factor that impacts the time required to achieve cost recovery from the EHR investments is clinic net revenue. With respect to how clinics fared financially upon adopting EHR systems, all but one of the primary care clinics in our study achieved an increase in clinic net revenue in the post-EHR period, as shown in Table 2.

Table 2. Percent changes in clinic revenues, net revenues, and clinician FTEs between the pre-EHR and post-EHR periods.

Clinic #	Percent change between the pre-EHR and post-EHR periods (in ascending order by percent change in number of clinician FTEs), %		
	In number of clinician FTEs	In clinic revenue	In clinic's net revenue
Clinic 1	-29	23	23
Clinic 2	-20	-28	22
Clinic 3	-14	27	4
Clinic 4	-2	29	26
Clinic 7	0	55	9
Clinic 5	0	50	63
Clinic 9	0	33	8
Clinic 10	0	31	28
Clinic 8	0	23	28
Clinic 11	0	19	16
Clinic 6	0	3	15
Clinic 12	0	-10	20
Clinic 13	0	-15	-30
Clinic 14	10	120	116
Clinic 15	47	223	227
Clinic 16	53	103	98
Clinic 17	329	603	845
Average	22	76	89

In addition to clinic net revenue, the sampled clinics showed, on average, positive increases in active patient count, clinician count, clinical support staff count, and clinic revenue in the post-EHR implementation period. These increases are summarized in Figure 1.

Percent increase in clinic net revenue between the pre-EHR and post-EHR periods showed a very strong positive correlation with percent increase in clinic revenue in the same periods (r=.99). Percent increase in clinic revenue showed a strong positive correlation with percent increase in the number of active patients (r=.87). It also showed a strong positive correlation with percent increase in the number of clinician FTEs, as well as with the number of clinical-support-staff FTEs (r=.96 and r=.97, respectively). These correlation coefficients (r values) are summarized in Figure 2.

Figure 1. Average percent changes in active patient count, clinician FTE count, clinical support staff FTE count, clinic revenue, and clinic net revenue between the pre-EHR and post-EHR periods.

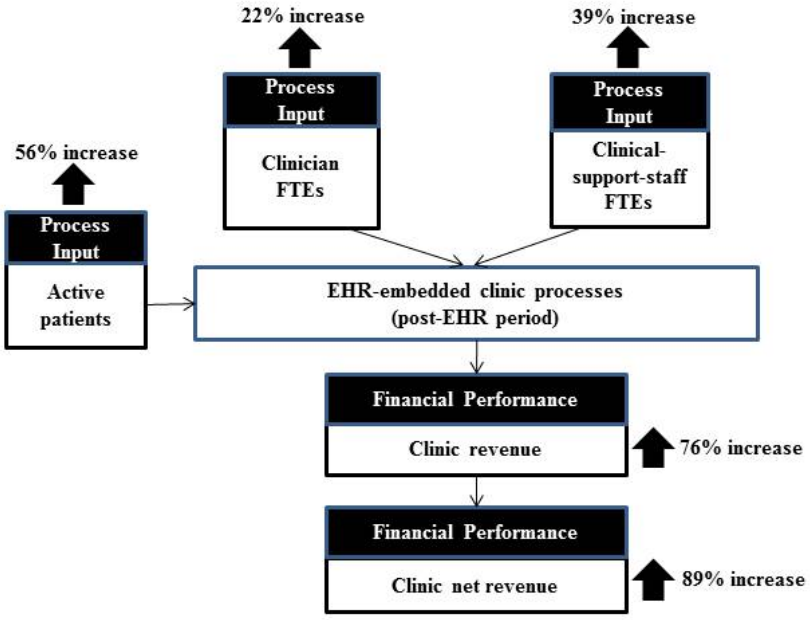
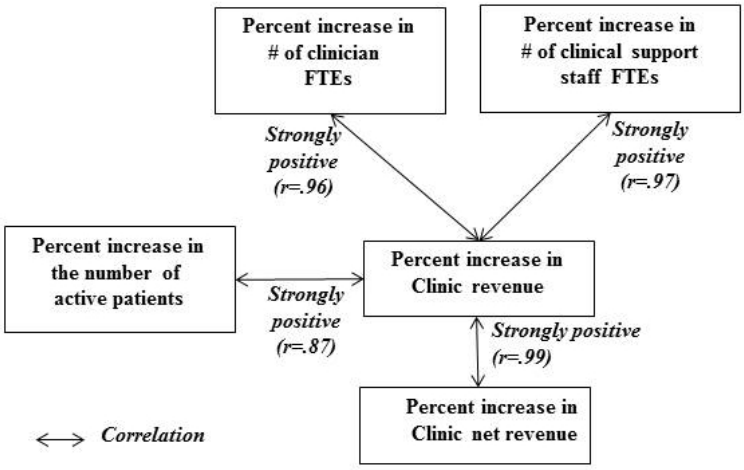


Figure 2. Correlations (r-values): clinic net revenue, clinic revenue, active patient count, clinician FTE count, and clinical support staff FTE count.



Percent Changes of Counts After Implementation—Not Linearly Proportional to One Another

Interestingly, the percent increases in active patient count, clinician FTE count, and clinical support staff FTE count are not linearly proportional to one another. An average active-patient-count increase of 56% was handled by an average 22% increase for clinician FTEs and an average 39% increase for clinical-support-staff FTEs. This finding suggests change in operational efficiency after EHR implementation, with respect to the active-patients-to-clinician-FTE ratio and the active-patients-to-clinical-support-staff-FTE ratio. The sampled clinics showed an average increase of 27% in the active-patients-to-clinician-FTE ratio and an average increase of 10% in the active-patients-to-clinical-support-staff-FTE ratio, as illustrated in Figure 3.

Percent increase in the number of active patients showed strong positive correlations with percent increases in active-patients-to-clinician-FTE ratio (r=.64) and in active-patients-to-clinical-support-staff-FTE ratio (r=.70), as shown in Figure 4.

These correlations, together with the nonlinear percent changes summarized in Figure 3, suggest that the increased efficiency in the post-EHR period contributed to a clinic's ability to accommodate the increased number of active patients.

Figure 3. Average percent changes in a clinic's operational efficiency and financial performance between the pre-EHR and post-EHR periods.

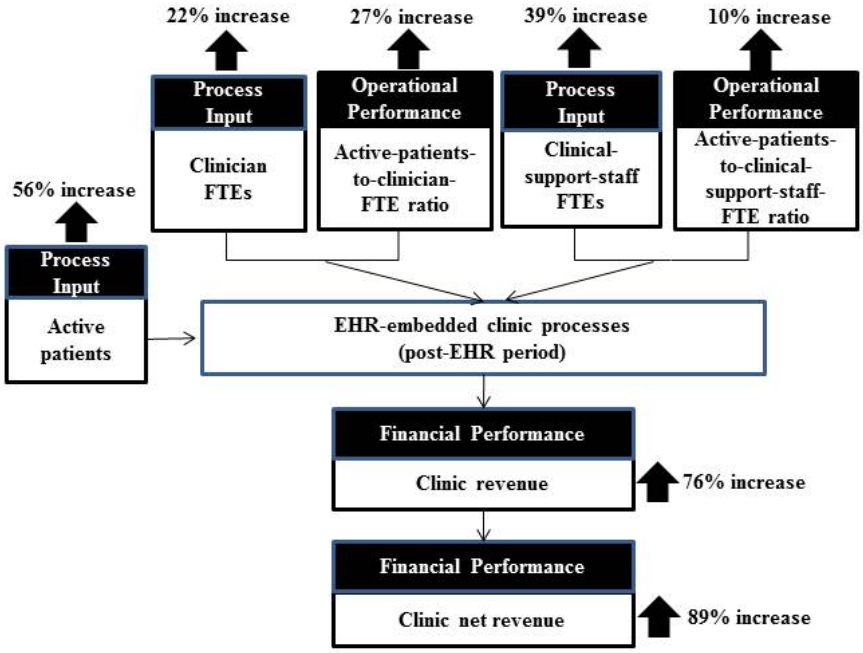
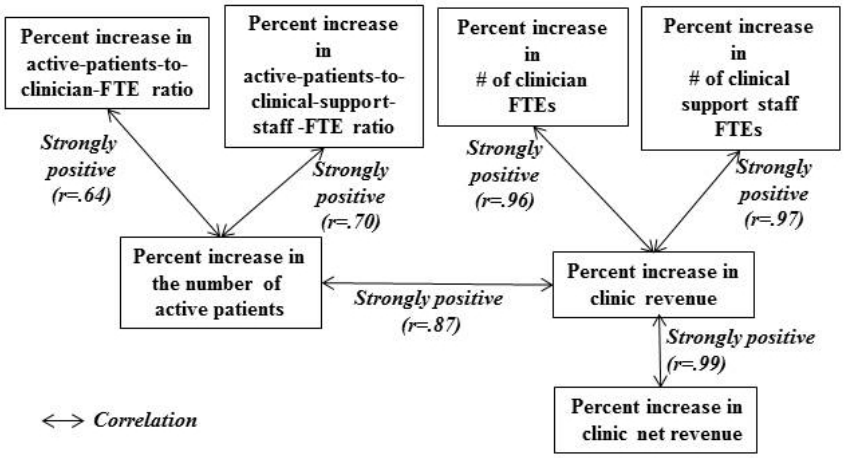


Figure 4. Correlations (r-values): clinic net revenue, clinic revenue, active patient count, clinician FTE count, clinical support staff FTE count, active-patients-to-clinician-FTE ratio, and active-patients-to-clinical-support-staff-FTE ratio.



Percent Changes in Number of Active Patients and Revenue After Implementation—Not Linearly Proportional to One Another

The percent increase in clinic revenue was also not linearly proportional to the percent increase in the number of active patients—an average increase of 76% versus an average increase of 56%, respectively. In addition to the increase in the number of active patients, there seem to be other factors that contributed to clinic revenue increase in the post-EHR period (detailed analysis on the impact of EHR on the sampled clinics' billing patterns and revenue management processes, required to identify the contributing elements of the greater than linear increase in clinic revenue over the increase in patient count, is beyond the scope of the study).

The study also finds that percent increase in clinic net revenue was not linearly proportional to percent increase in clinic revenue. The average additional 13% increase in clinic net revenue (89%, which is 13% above the clinic average revenue of 76%) is attributable to the enhanced operational efficiency in the post-EHR period, which suggests the relative cost-savings effect after the EHR implementation.

Sign Test Results

We further tested the financial and operational impact of EHR in the post-EHR period, in order to assess the degree to which these findings could be extended to the population of clinics implementing EHR. The sign test, as opposed to *t* test, was adopted because the sample size was less than 30 and because the distributions shown in the data were not normal, with a high degree of skewness in most cases. The sign test results shown in Table 3 suggest, with a 95% confidence level, that the median percent change in clinic net revenue between the pre-EHR and post-EHR periods is positive in the population of the primary care clinics implementing EHR (sign test M=7.5 with $P<.001$).

Table 3. Summary of statistical analysis of change in a clinic's operational efficiency and financial performance between between the pre-EHR and post-EHR periods

Percent changes between the pre-EHR and post-EHR periods	Average	SD	Median	M	Sign test, *P* value
Percent change in clinic net revenue	89%	203%	23%	7.5	<.001
Percent change in the number of active patients	56%	119%	10%	5.0	.006
Percent change in active-patients-to-clinician-FTE ratio	27%	53%	9%	6.5	<.001
Percent change in the number of clinician FTEs	22%	82%	0%	0.0	1.00
Percent change in the number of clinical support staff FTEs	37%	75%	0%	3.0	.07
Percent change in active-patients-to-clinical-support-staff-FTE ratio	10%	29%	4%	2.5	.277

The same conclusions can be made for the median percent changes in the active-patients-to-clinician-FTE ratio and in the number of active patients in the same periods (M=6.5 and M=5, respectively). However, for the median percent changes with respect to the number of clinician FTEs, the number of clinical support staff FTEs, and the active-patients-to-clinical-support-staff-FTE ratio, we could not reject with a 95% confidence level the null hypothesis of no change after EHR implementation.

The correlation coefficients shown in Figure 4 and sign test results summarized in Table 4 suggest that the increase in the active patient count may not be the only factor that contributed to an average break-even point of 10 months upon EHR implementation. Percent increases in the number of active patients, in the active-patients-to-physician-FTE ratio, and in clinic net revenue appear to be positively associated with the EHR implementation, likely contributing substantially to an average break-even point of 10 months.

Analysis of Variance in Realizing Financial Performance—Key Factors

Study participants reported improvements in their ability to manage patient information after the implementation of EHR systems, citing improved ability to manage results such as obtaining test results from laboratories and following the results of an investigation over time (64%, 11/17 clinics). Respondents also reported an improved ability to seek out specific information from patient records (57%, 10/17 clinics), and access complete, up-to-date patient charts and review patient problems (43%, 6/15 clinics). See Table 4 for key EHR impacts expressed by study participants.

Table 4. Impact of EHR on clinic practices identified by study participants.

Categories	Participant comments
A. Impact of EHR on a clinic's ability to manage results	"We receive results electronically and can graph them; graphs help 'engage' the patient."
	"Direct to physician lab results has very positive effect on physician efficiency and patient care."
B. Impact of EHR on a clinic's ability to seek out specific information from patient records	"Complete chart is always available, anywhere which affects patient safety and means better care."
	"Integration of information for referral requests is a great benefit."
	"Billing codes are up-to-date. (And) billing is automatic by the doctor inside encounter note, which simplifies billing and is easier to manage reconciliation. No missed billing opportunities."
C. Impact of EHR on a clinic's ability to prepare patient encounter	"Review of patient information prior to encounter is greatly facilitated."
	"Easier to prepare for encounter; maintenance of problem list /summary is much easier"
	"Immediate access to patient information—no lost files."

Some primary care clinics did better than others in using EHR and achieving faster break-even from EHR investment, which can be observed in Tables 1 and 3. To gain insights into key differences between those clinics that were highly successful and those less successful in realizing a positive ROI from EHR, we conducted regression analysis on break-even point as the outcome variable. We used the codes identified through the coding process of the qualitative data as a part of the explanatory variables to estimate their relationships with the outcome variable "break-even point". As summarized in Table 5, the regression analysis suggests four statistically significant factors impacting the return on EHR investment, that is, the time required to achieve cost recovery from an investment in EHR.

Table 5. Significant linear regression results of the outcome "break-even point" with explanatory variables (break-even point was log-transformed to approach a Normal distribution).

Explanatory variable	Variable values	Regression coefficient	Standard error of coefficient	*P* value	r^2
(a) Age of EHR: Months between Jan 1, 2013, and EHR implementation start date	Number of months	0.03	0.01	.049	.64
(b) e-Prescriptions complying with national standards	0 (No) to 1 (Yes)	-1.32	0.34	.006	.50
(c) Extent to which EHR complies to national standards	Continuous (from 0 to 10)	-0.19	0.07	.038	.54
(d) Process change: Use of flow sheets	0 (No) to 1 (Yes)	-1.29	0.46	.022	.68

Note that in Table 5, the regression coefficient of an explanatory variable with a negative value indicates faster recovery of the EHR investment (ie, a shorter time required to achieve cost recovery from an EHR investment), while a positive value implies slower recovery of the EHR investment (ie, a longer time required to achieve cost recovery from an EHR investment).

Age of Electronic Health Record Systems

The first result to note in item (a) of Table 5 is that older EHR implementations, in particular those implemented in 2004-2005, were slower to recover their investment, even though they still achieved a break-even point. This result suggests that the newer the EHR, the sooner a positive ROI can be achieved. The earlier EHR systems used by these clinics were less user-friendly and required longer training cycles for the users, which may explain why clinics with these earlier systems took longer to recoup their financial investment.

Compliance With National Standards

The second and third results, shown in items (b) and (c) of Table 5, suggest a positive link between the ROI indicator and the compliance with national standards such as codes representing prescription drugs. There was an improvement in clinics' compliance with national standards and ability to comply with evidence-based medicine. This improvement was related to the age of the EHR system used by the clinics. Newer EHR implementations may be more likely to comply with national standards, given that the newer EHRs are likely to support national standards better.

Use of Flow Sheets and Ability to Manage Patient Information

Finally, clinics using EHR flow sheets scored consistently better times to break even, shown in item (d) of Table 5. Clinics reported the use of flow sheets, or structured data collection forms, as a mechanism for compliance to evidence-based medicine. The use of flow sheets in EHRs provides advanced features such as those related to the automatic maintenance of

patient problem lists and pharmacological profiles. These enhanced features contribute directly to the physician's efficiency by eliminating the time that would otherwise be spent manually maintaining these lists—a task that can be time-consuming, highly repetitive, and labor-intensive to maintain with consistency in a paper-based environment. The availability of up-to-date lists makes patient encounter preparation easier and more rapid, as the necessary information is available at a glance.

Analysis of Electronic Health Record Functionalities Used in Primary Care Clinics

Our study finds that despite the limited use of EHR functionalities and limited interoperability, the sampled clinics achieved overall positive operational and financial performance. Table 6 summarizes the data we gathered on EHR functionalities, frequency of use, and ease of use.

Most frequently and routinely used EHR functionalities were related to medication management. Health information exchange and patient engagement portal functionalities saw no significant use (the investigation of why these functionalities were not used is beyond the scope of this study).

Respondents stressed that it typically takes a few months to understand any particular EHR function sufficiently to effectively introduce it in their clinical practices. This finding, coupled with the finding that despite the limited use of EHR functionalities the clinics achieved overall positive improvement in operational and financial performance in the post-EHR period, suggests that a clinic's ability to embed particular EHR functionalities in their workflow and make use of these functionalities in their day-to-day clinical practices is of more importance in realizing a positive ROI from EHR implementation than implementing an EHR software package with the maximum number of features and functionalities.

Table 6. EHR functionalities and utilization reported during the study period.

EHR functionalities	% of clinics answering in the affirmative
User Interface: Does the EHR system currently in use at this clinic have any of the following user interface technologies? (N=17)	
Alternative presentation formats for clinical information	100.0
Support for guideline-based data collection and treatment	94.1
Support for multiple platform access	88.2
Support for context sensitive alerts, warnings, and guidance	70.6
Clinical notes capture in narrative form	23.5
Listing functionality: With the EHR system you currently have, how easy is it for you (or staff in your practice) to generate the following information about your patients? (N=17)	
List of all medications taken by an individual patient	100.0
Provide patients with clinical summaries for each visit	88.2
List of all laboratory results for an individual patient	88.2
List of patients by diagnosis (eg, diabetes or cancer)	82.4
List of patients who are due or overdue for tests or preventive care	76.5
List of all patients taking a particular medication	76.5
List of patients by laboratory result (eg, HbA1C>9.0)	52.3
Reminder functionality: Are the tasks routinely performed for patients at your site using EHR? (N=17)	
Clinicians receive a reminder for guideline-based interventions and/or screening tests	58.8
Clinicians receive an alert or prompt to provide patients with test results	41.2
Patients are sent reminder notices when it is time for regular preventive or follow-up care	35.3
All laboratory tests ordered are tracked until results reach clinicians	29.4
Does your site and the clinicians that practice in your site use the EHR system to facilitate any of the following workflow activities (N=16)	
Electronic prescribing of medication	93.3
Electronic prompts about a potential problem with drug dose or drug interaction	87.5
Electronic receipt of laboratory results integrated into the EHR system (not scanned)	62.5
Electronic ordering of laboratory tests	43.8
Electronic referring to specialists	37.5
electronic transferring of prescriptions to a pharmacy	6.7
Health information exchange functionalities: Can you electronically exchange the following with any doctors outside your practice? (N=16)	
Electronic exchange outside practice: patient clinical summaries	25.0
Electronic exchange outside practice: laboratory and diagnostic tests	18.8
Patient engagement functionality: Please indicate whether the EHR system in use at your site allows patients to… (N=17)	
Access alcohol consumption advice online	11.8
Access advice for informal caregivers online	11.8
Email about a medical question or concern	11.8
Access dietary advice online	11.8
Access advice on physical activity online	11.8
Access advice on self-management of chronic conditions online	11.8
Access smoking cessation advice online	11.8
Request appointments online	5.9
View a list of medications (current and past) online	5.9
View other components of their chart (current and past) online	0.0
View medical imaging results (current and past) online	0.0

EHR functionalities	% of clinics answering in the affirmative
Request refills for prescriptions online	0.0
View test results (current and past) online	0.0
Interoperability: Were any of the following INTEROPERABILITY technologies implemented in the EHR system currently in use at this site? (N=17)	
Diagnoses are coded using international standards	94.1
Medications and pharmacological profiles are coded to national standards	82.4
Patient records are supported by standards-based data migration technology	50.0
ePrescriptions comply with national standards	52.9
Patient Identifier is based on national or jurisdictional standard	58.8
Patient Identifier is supported by aliasing technology to achieve positive ID across systems	37.5
Findings are coded using international standards	58.8
Communications with other clinics and institutions use international standards	31.3
Investigations, referrals, and imaging requests make use of order tracking technology	35.3

Discussion

Principal Findings

This study aimed to complement current insights into the cost recovery concerns related to EHR investments by considering the research question "When implemented in primary care practices, what will be the ROI from EHR?". The study finds that primary care clinics can realize a positive ROI from the implementation of EHR. Our analysis offers evidence that the increases in net revenue, in the active-patients-to-clinician-FTE ratio, and in the number of active patients are positively associated with the EHR implementation, likely contributing substantially to an average break-even point of 10 months.

In addition, the analysis conducted to understand the variances in financial and operational performance among the sampled clinics provides insights into key differences between those clinics that were highly successful and those less successful in realizing a positive ROI from EHR. Some clinics seem to be more innovative than others in using EHR in their practices to achieve significantly better operational and financial results. The analysis suggests that a clinic's ability to take advantage of EHR to support process changes has a significant effect on the time required to achieve cost recovery from an investment in EHR. In particular, the clinics that were successful in realizing faster time to break even were better at using EHR in workflow areas involving patient information—such as maintaining patient problem lists, managing test results, and complying with national coding standards, all of which make patient encounter preparation easier and more rapid. We also find that the clinics achieved positive financial performance, even though not all EHR functionalities were used. The alignment of EHR functionalities with clinic workflow plays an important role in achieving positive operational and financial results with EHR. Identified as particularly important EHR-product improvements that would ease adoption of workflow changes are automations that assist clinicians, clinical support staff, and administrative staff both in the overall management of the practice and within the patient encounter, as well as consistent and comprehensive compliance with national standards such as national drug coding standards.

Implications for Practitioners and Managers in Primary Care

The knowledge gained from this ROI study on EHR is important to practicing primary care physicians who are concerned about how they will fare financially upon investing in EHR, as they face ever increasing pressure to transition from their paper-based records to electronic systems. This study provides evidence to practitioners in primary care that investment in EHR can be a sound decision with a reasonable cost recovery time frame, while providing immediate opportunities for increased operational efficiency and the potential for further improvements in clinic performance and benefits realization from EHR. Practitioners in primary care who are considering the investment in EHR should note the important relationship between EHR functionality, clinic workflow change, and a positive ROI from EHR implementation. Positive ROI does not happen automatically upon implementing an EHR package, and a clinic's ability to leverage EHR for process changes plays a role in achieving a positive ROI.

Implications for Policy Makers

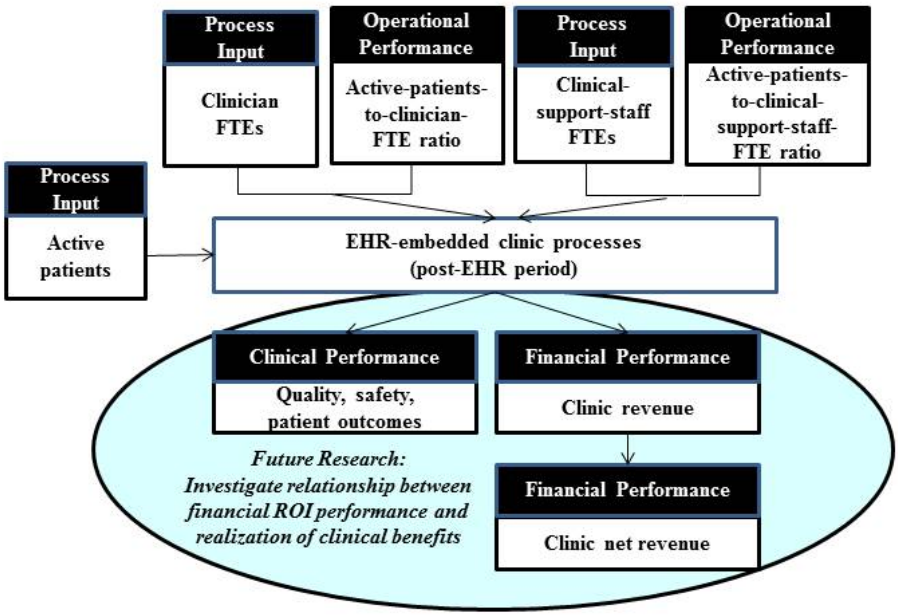
This study's finding on increased active patient count and clinic operational efficiency after the EHR implementation, in particular with respect to improvement in the active-patients-to-clinician-FTE ratio, offers the possibility that EHR can play a role in addressing the shortage in family physicians. As primary care clinics implement EHR systems and discover better ways to take advantage of EHR in their practices, a key question will be how to incorporate such learnings and deliver enhanced EHR products back into the clinics to realize the full potential of EHR. Policies that enable the establishment of a closed-loop feedback mechanism between EHR vendors and health care providers could facilitate targeted enhancements to EHR systems. In addition, policies that provide support to help primary care practices successfully make EHR-enabled changes, such as support of workflow optimization with an EHR system that would ease adoption, could not only

facilitate the realization of positive ROI but also help address the shortage in family physicians.

Future Research

Some of the factors identified in this research as key factors impacting the realization of a positive ROI from EHR implementation, such as improved access to up-to-date patient charts and improved ability to obtain test results from laboratories and follow the results of an investigation over time, have implications to quality of care and patient safety. Thus, future research will be to investigate the relationship between financial ROI and realization of clinical benefits of EHR such as quality, safety, and patient outcomes, as depicted in Figure 5. Other research should include a study to identify best practices for implementing and using EHR, with concrete examples of success factors and failure factors as well as ways to tailor these best practices relevant to particular clinic situations. In addition, panel analysis, which deals with two-dimensional panel data (cross sectional and times series) [35], can be conducted with the cohort of primary care clinics to understand the effect of learning curve on a clinic's ability to realize positive ROI and non-financial, clinical benefits from EHR implementation. Knowledge gained from such studies could facilitate EHR adoption and subsequent benefits realization in primary care practices.

Figure 5. Future research: investigate the relationship between return on EHR investment and clinical benefits realization from EHR implementation.



Limitations

The principal limitation of this study is that the number of primary care clinics examined was limited, due mainly to time constraints of clinics to participate in the study and scarcity of suitably detailed operational and financial data necessary for ROI calculation. For the clinics recruited to the study, the most limiting factor was that of collecting a complete picture of the cost and benefits needed to assess an ROI from EHR implementation. This was due mainly to the absence of standardized financial and business-case approaches to the governance of these independent organizations. The insights gained from the participants in our study, however, provide salient insights into the impact of EHR investment to facilitate the EHR adoption across practicing primary care physicians, with information on time required to achieve cost recovery from an EHR investment and on principal factors impacting cost-recovery performance.

Acknowledgments

The authors wish to acknowledge the valuable contributions of Liette Lapointe, PhD, Isabelle Vedele, PhD, John Hughes, MD, Raymond Simkus, MD, and Susan Law, PhD, in all phases of this study.

This study was funded by Canada Health Infoway Inc, an independent, not-for-profit corporation established in Canada to accelerate the adoption of electronic health records and related technologies on a pan-Canadian basis. The opinions, results, and conclusions reported in this manuscript are those of the authors. No endorsement by Canada Health Infoway is intended or should be inferred.

Conflicts of Interest

None declared.

References

1. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, et al. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Heath Affairs 2005 Sep;24(5):1103-1117 [FREE Full text] [doi: 10.1377/hlthaff.24.5.1103]
2. Blumenthal D. Stimulating the adoption of health information technology. N Engl J Med 2009 Apr 9;360(15):1477-1479 [FREE Full text] [doi: 10.1056/NEJMp0901592]
3. Bates DW. Getting in step: electronic health records and their role in care coordination. J Gen Intern Med 2010 Mar;25(3):174-176 [FREE Full text] [doi: 10.1007/s11606-010-1252-x] [Medline: 20127195]
4. Saleem JJ, Flanagan ME, Wilck NR, Demetriades J, Doebbeling BN. The next-generation electronic health record: perspectives of key leaders from the US Department of Veterans Affairs. J Am Med Inform Assoc 2013 Jun;20(e1):e175-e177 [FREE Full text] [doi: 10.1136/amiajnl-2013-001748] [Medline: 23599227]
5. Black AD, Car J, Pagliari C, Anandan C, Cresswell K, Bokun T, et al. The impact of eHealth on the Quality and Safety of Health Care: a Systematic Overview. PLoS Med 2011;8(1) [FREE Full text] [doi: 10.1371/journal.pmed.1000387]
6. Holroyd-Leduc JM, Lorenzetti D, Straus SE, Sykes L, Quan H. The impact of the electronic medical record on structure, process, and outcomes within primary care: a systematic review of the evidence. Journal of the American Medical Informatics Association 2011 [FREE Full text] [doi: 10.1136/amiajnl-2010-000019]
7. Gans D, Kralewski J, Hammons T, Dowd B. Medical groups' adoption of electronic health records and information systems. Health Affairs 2005;24(5):1323-1333 [FREE Full text] [doi: 10.1377/hlthaff.24.5.1323] [Medline: 16162580]
8. DesRoches CM, Campbell EG, Rao SR, Donelan K, Ferris TG, Jha A, et al. Electronic health records in ambulatory care--a national survey of physicians. N Engl J Med 2008 Jul 3;359(1):50-60 [FREE Full text] [doi: 10.1056/NEJMsa0802005] [Medline: 18565855]
9. El-Kareh R, Gandhi TK, Poon EG, Newmark LP, Ungar J, Lipsitz S, et al. Trends in primary care clinician perceptions of a new electronic health record. Journal of General Internal Medicine 2009 Jan;24(4):464-468 [FREE Full text] [doi: 10.1007/s11606-009-0906-z]
10. Bassi J, Lau F, Lesperance M. Perceived impact of electronic medical records in physician office practices: a review of survey-based research. Interact J Med Res 2012;1(2):e3 [FREE Full text] [doi: 10.2196/ijmr.2113] [Medline: 23611832]
11. Keshavjee K, Bosomworth J, Copen J, Lai J, Kucukyazici B, Lilani R, et al. Best practices in EMR implementation: a systematic review. 2006 Presented at: AMIA Annu Symp Proc; 2006; USA p. 982 URL: http://www.infoclin.ca/assets/7e474_best%20practices%20in%20emr%20implementation%20-%20july,%202006.pdf
12. Ilie V, Van Slyke C, Parikh MA, Courtney JF. Paper Versus Electronic Medical Records: The Effects of Access on Physicians' Decisions to Use Complex Information Technologies. Decision Sciences 2009;40(2):213-241 [FREE Full text] [doi: 10.1111/j.1540-5915.2009.00227.x]
13. Valdes I, Kibbe DC, Tolleson G, Kunik ME, Petersen LA. Informatics in primary care. 2004. Barriers to proliferation of electronic medical records URL: http://www.ingentaconnect.com/content/bcs/ipc/2004/00000012/00000001/art00002 [WebCite Cache ID 6Ra6q7TOp]
14. Archer N, Cocosila M. A comparison of physician pre-adoption and adoption views on electronic health records in Canadian medical practices. Journal of medical Internet research 2011;13(3). [doi: 10.2196/jmir.1726]
15. Terry A, Thorpe C, Giles G, Brown J, Harris S, Reid G, et al. Implementing electronic health records: Key factors in primary care. Can Fam Physician 2008 May;54(5):730-736 [FREE Full text] [Medline: 18474707]
16. Ludwick DA, Doucette J. Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries. Int J Med Inform 2009 Jan;78(1):22-31. [doi: 10.1016/j.ijmedinf.2008.06.005] [Medline: 18644745]
17. Boonstra A, Broekhuis M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. BMC Health Serv Res 2010;10:231 [FREE Full text] [doi: 10.1186/1472-6963-10-231] [Medline: 20691097]

18. Simon SR, Kaushal R, Cleary PD, Jenter CA, Volk LA, Poon EG, et al. Correlates of electronic health record adoption in office practices: a statewide survey. J Am Med Inform Assoc 2007;14(1):110-117 [FREE Full text] [doi: 10.1197/jamia.M2187] [Medline: 17068351]
19. Kemper AR, Uren RL, Clark SJ. Adoption of electronic health records in primary care pediatric practices. Pediatrics 2006;118(1) [FREE Full text] [doi: 10.1542/peds.2005-3000]
20. Rozenblum R, Jang Y, Zimlichman E, Zalzberg C, Tamblyn M, Buckeridge D, et al. A qualitative study of Canada's experience with the implementation of electronic health information technology. Canada Med Assoc J 2011;183(5) [FREE Full text] [doi: 10.1503/cmaj.100856]
21. Pifer EA, Smith S, Keever GW. EMR to the rescue. An ambulatory care pilot project shows that data sharing equals cost shaving. Healthc Inform 2001 Feb;18(2):111-114. [Medline: 11225061]
22. Wang SJ, Middleton B, Prosser LA, Bardon CG, Spurr CD, Carchidi PJ, et al. A cost-benefit analysis of electronic medical records in primary care. Am J Med 2003 Apr 1;114(5):397-403. [Medline: 12714130]
23. Miller RH, West C, Brown TM, Sim I, Ganchoff C. The value of electronic health records in solo or small group practices. Health Aff (Millwood) 2005;24(5):1127-1137 [FREE Full text] [doi: 10.1377/hlthaff.24.5.1127] [Medline: 16162555]
24. Grieger DL, Cohen SH, Krusch DA. A pilot study to document the return on investment for implementing an ambulatory electronic health record at an academic medical center. J Am Coll Surg 2007 Jul;205(1):89-96. [doi: 10.1016/j.jamcollsurg.2007.02.074] [Medline: 17617337]
25. Adler-Milstein J, Green CE, Bates DW. A survey analysis suggests that electronic health records will yield revenue gains for some practices and losses for many. Health Aff (Millwood) 2013 Mar;32(3):562-570. [doi: 10.1377/hlthaff.2012.0306] [Medline: 23459736]
26. Sahin D, Yaffe MJ, Sussman T, McCusker J. A mixed studies literature review of family physicians' participation in research. Fam Med 2014;46(7):503-514 [FREE Full text] [Medline: 25058542]
27. Creswell J, Plano Clark V. Designing and conducting mixed methods research. Thousand Oaks: Sage Publications; 2010.
28. Lærum H, Faxvaag A. Task-oriented evaluation of electronic medical records systems: development and validation of a questionnaire for physicians. BMC medical informatics and decision making 2004;4(1) [FREE Full text] [doi: 10.1186/1472-6947-4-1]
29. Joos D, Chen Q, Jirjis J, Johnson KB. An electronic medical record in primary care: impact on satisfaction, work efficiency and clinic processes. 2006 Presented at: AMIA Annual Symposium; 2006; USA p. 394 URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839545/#__ffn_sectitle
30. Aaronson JW, Murphy-Cullen CL, Chop WM, Frey RD. Electronic medical records: the family practice resident perspective. Fam Med 2001 Feb;33(2):128-132. [Medline: 11271741]
31. Audet AM, Doty MM, Shamasdin J, Schoenbaum SC. Measure, learn, and improve: physicians' involvement in quality improvement. Health Aff (Millwood) 2005;24(3):843-853 [FREE Full text] [doi: 10.1377/hlthaff.24.3.843] [Medline: 15886180]
32. Otieno OG, Toyama H, Asonuma M, Kanai-Pak M, Naitoh K. Nurses' views on the use, quality and user satisfaction with electronic medical records: questionnaire development. J Adv Nurs 2007 Oct;60(2):209-219. [doi: 10.1111/j.1365-2648.2007.04384.x] [Medline: 17877568]
33. Sittig DF, Kuperman GJ, Fiskio J. Evaluating physician satisfaction regarding user interactions with an electronic medical record system. 1999 Presented at: AMIA Symposium; 1999; USA p. 400.
34. Glaser BG, Strauss AL. The discovery of grounded theory: Strategies for qualitative research. In: Aldine Transaction. Piscataway, NJ: Aldine Transaction; Dec 1999.
35. Hsiao C. Analysis of Panel Data, 2nd edition. Cambridge: Cambridge University Press; 2003.

Abbreviations

EHR: electronic health records
FTE: full time equivalent
IT: information technology
ROI: return on investment

Edited by G Eysenbach; submitted 20.06.14; peer-reviewed by J Tan, G Nasi; comments to author 23.07.14; revised version received 05.08.14; accepted 11.09.14; published 29.09.14.

Please cite as:
Jang Y, Lortie MA, Sanche S
Return on Investment in Electronic Health Records in Primary Care Practices: A Mixed-Methods Study
JMIR Med Inform 2014;2(2):e25
URL: http://medinform.jmir.org/2014/2/e25/
doi:10.2196/medinform.3631
PMID:25600508

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org