JMIR Medical Informatics

Impact Factor (2023): 3.1
Volume 2 (2014), Issue 1 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Original Papers

Viewpoints

Reviews

Original Paper

Designing an Algorithm to Preserve Privacy for Medical Record Linkage With Error-Prone Data

Doyel Pal[1]; Tingting Chen[1], PhD Comp Sc; Sheng Zhong[2,3], PhD Comp Sc; Praveen Khethavath[1]

[1]Computer Science Department, Oklahoma State University, Stillwater, OK, United States

[2]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

[3]Department of Computer Science and Technology, Nanjing University, Nanjing, China

Corresponding Author:
Sheng Zhong, PhD Comp Sc
State Key Laboratory for Novel Software Technology
Nanjing University
Department of Computer Science and Technology
Nanjing University
Nanjing, 210023
China
Phone: 86 25 89680686
Fax: 86 25 89683216
Email: sheng.zhong@gmail.com

Abstract

Background: Linking medical records across different medical service providers is important to the enhancement of health care quality and public health surveillance. In records linkage, protecting the patients' privacy is a primary requirement. In real-world health care databases, records may well contain errors due to various reasons such as typos. Linking the error-prone data and preserving data privacy at the same time are very difficult. Existing privacy preserving solutions for this problem are only restricted to textual data.

Objective: To enable different medical service providers to link their error-prone data in a private way, our aim was to provide a holistic solution by designing and developing a medical record linkage system for medical service providers.

Methods: To initiate a record linkage, one provider selects one of its collaborators in the Connection Management Module, chooses some attributes of the database to be matched, and establishes the connection with the collaborator after the negotiation. In the Data Matching Module, for error-free data, our solution offered two different choices for cryptographic schemes. For error-prone numerical data, we proposed a newly designed privacy preserving linking algorithm named the Error-Tolerant Linking Algorithm, that allows the error-prone data to be correctly matched if the distance between the two records is below a threshold.

Results: We designed and developed a comprehensive and user-friendly software system that provides privacy preserving record linkage functions for medical service providers, which meets the regulation of Health Insurance Portability and Accountability Act. It does not require a third party and it is secure in that neither entity can learn the records in the other's database. Moreover, our novel Error-Tolerant Linking Algorithm implemented in this software can work well with error-prone numerical data. We theoretically proved the correctness and security of our Error-Tolerant Linking Algorithm. We have also fully implemented the software. The experimental results showed that it is reliable and efficient. The design of our software is open so that the existing textual matching methods can be easily integrated into the system.

Conclusions: Designing algorithms to enable medical records linkage for error-prone numerical data and protect data privacy at the same time is difficult. Our proposed solution does not need a trusted third party and is secure in that in the linking process, neither entity can learn the records in the other's database.

(JMIR Med Inform 2014;2(1):e2) doi:10.2196/medinform.3090

KEYWORDS

privacy; medical record linkage; error-prone data

Introduction

Electronic Patient Records

With the popularity of electronic patient records and the expanded use of medical information systems [1], nowadays many different health care providers store medical records of patients electronically. In many cases different health care providers hold the same patient's data. To enhance the quality of health care treatment, for example, in regional health information networks, often it is required to gather information about the same patient from different providers [2]. In order to identify whether a particular patient's information is held by more than one health care provider or not, a matching technique is used on the key attributes of the patient's demographic information [2]. As another example, public health surveillance often requires linking patient records from different health care providers [1]. In order to monitor the quality of health care treatment provided in a region and to analyze a patient's medication interaction, it is very helpful to collect correlated data from different sources [3], (eg, clinics, pharmacy, laboratory, and health care providers).

Keeping Patient Information Secure and Private

With the increasing needs of keeping and linking electronic patient records, it is very challenging to maintain security and preserve privacy. Under the regulations of the Health Insurance Portability and Accountability Act (HIPAA) [4], preserving patient's privacy is important in linking the patient's data. As medical databases contain different identifiers of a patient (eg, patient's name, surname, date of birth, Social Security Number-SSN, contact number, address, etc), using these identifiers in their actual form for linking purpose violates privacy. Moreover, due to privacy, security, and business concerns, different health care providers may not be willing to reveal their health data information other than the linking result to the other provider. Among existing research works, Shapiro et al and Vest [1,2] illustrated some approaches toward health information exchange. One obvious approach is to link data using the identifiers in encrypted format [5-8]. An elegant approach to encrypt identifiers is using one-way hash functions as in Quantin et al and Quantin et al [7,8]. To ensure security, these methods are based on the irreversible transformation property of one-way hash functions on identification data. These methods are vulnerable to some common cryptographic attacks. In Quantin et al [8], the authors proposed a computerized hash encoding and anonymous record linkage procedure on medical information. To consolidate security against dictionary attack, Quantin et al [8] used two pads, one for the sender and the information and the other one for the recipient. Some other approaches have been proposed regarding privacy preserving medical records linkage algorithms [9,10]. A trusted third party has been used in Churches and Christen [10], to make the algorithm more secure. Here each party is involved in computing the set of bigrams for each string. Each party exchanges the power set of encrypted bigrams with the trusted third party and then string similarity is performed using the Dice coefficient. However, these approaches usually have high false negative rates. Using indirect pseudonym identifiers [9], besides giving the patients control over what information is revealed about them, an architecture has been proposed to link medical records.

Some algorithms on privacy preserving data matching are proposed in database and data mining research fields. In Lindell and Pinkas [11], the authors have proposed a solution where two parties can run a data mining algorithm on the union of their own confidential databases, without revealing any unnecessary information. In this particular solution, the authors focused on the problem of decision tree learning, as the input sizes of data mining algorithms are huge and the data mining algorithms themselves are very complex. At each party's end, this method uses a computation of the same order as computing the Iterative Dichotomiser 3 algorithm on its own databases. It combines the result using cryptographic tools. Some solutions of privacy preserving record linkage are based on the perturbed information. For example, Agrawal and Srikant [12] used a randomizing function such as the Gaussian function or uniform perturbations to perturb the sensitive data and build a decision tree classifier from these perturbed data. This solution offers a reconstruction procedure to accurately estimate the original data value distribution. The cryptographic technique, which relies on the secure multi-party computation (SMC) Protocol [13], computes functions over private inputs. Scannapieco et al [14] proposed a more efficient protocol based on cryptographic techniques, which preserves privacy of database schemas. Here, the authors consider the scenario that two parties want to link their data in string format and can have different schemas. They propose a protocol that consists of data matching and schema matching protocol. The protocol builds an embedding space and two parties embed their data strings using a variant of Sparse Map [15] ensuring the contractiveness of the embedding. A semihonest third party collects the embedded strings from the two parties and computes the similarities. Whereas Agrawal and Srikant [12] concentrated on exact matching, Scannapieco et al [14] focused on approximate matching. A hybrid method that combines both the data perturbation and cryptographic techniques is presented in Inan et al [16]. The basic idea of this method is to first classify the data into two classes as matches and mismatches, and then apply the general SMC protocol to compute the distance for the records in the matches' class. A querying party is introduced to provide the classifier that determines matching record pairs. The problem with this method is that general SMC protocols are costly to use in practice. In Scannapieco et al and Inan et al [14,16], the proposed solutions used a trusted third party, which is a major issue since the Web-based trusted third party may not be a good choice to link records in a privacy preserving way.

Some recent works [17-20] focused on security and privacy in biological services and in medical data. By secure encapsulation and publishing of bioinformatics software in a cloud computing environment, Zhang et al [17] have derived a prototype system of the biological cloud. While they worked on only biological services and focused on achieving a prototype system of the biological cloud, our solution works on different databases and concentrates on linking these different databases in a privacy preserving manner. In Gkoulalas-Divanis and Loukides [18], the authors have discussed medical data sharing by preserving privacy and data utility. Here they have given a clear picture

about the data that generally is used for data sharing purposes, different techniques for privacy preserving data sharing, and different types of threats. A new algorithm has been proposed in Mohammed et al [20] for heterogeneous health data sharing in privacy preserving manner. The proposed algorithm considers health data containing both relational and set-valued data and accomplishes "element-of" differential privacy. In Gkoulalas-Divanis and Loukides [18], the authors have discussed different types of medical data, for example, demographics, clinical information, text, and genomic information, and Mohammed et al [20] worked on heterogeneous medical data. In our solution, instead of different categories of medical data and heterogeneous health data, we concentrate on textual and numeric data and further categorize them into error-free data and error-prone data. Kum et al [19] focused on privacy preserving interactive record linkage. The authors have given a solution by proposing a computer-based third party record linkage platform, Secure Decoupled Linkage. The proposed solution decouples the data, obfuscates it, and shares minimum information via encryption, chaffing, and recoding respectively, to ensure the protection against attribute disclosure. A new computer-based third party record linkage platform has been proposed in Kum et al [19], but our proposed solution does not need a trusted third party.

However, when we consider the real life scenario, it is possible that existing works might not meet all the requirements of medical record linkage all the time in practice. For instance, earlier researches [5] on data record linkage (ie, sending identifiers in encrypted format does not allow any kind of error in identifiers) may happen frequently in real cases. Spelling mistakes and typographical errors are very common in databases. Some researches [21-24] have been done toward the error-prone data and on the missing data. In Weber et al [24], the authors have proposed a solution to build cross-site records and link data for a particular patient as he/ she moves between participating sites. They considered the hypothesis that most variation in names occurs after the first two letters; this, along with the date of birth, is one of the most reliable attributes. Out of this consideration they generated the composite identifier based on the real identifiers in such a way that the possibility of identifying a common patient is maximized. This composite identifier is the hashed string of the first two letters of the patient's first name and last name, plus their date of birth. Considering this composite identifier, they have shown that it has a higher sensitivity rate compared to other identifiers (eg, SSN and identifier based on patient's full name and date of birth).

Most of the existing algorithms for error-prone data are concentrated on textual data. They are very useful for linking records for any customer identifying information. Some approaches toward error-prone data in privacy preserving record linkage have been proposed. One of these proposed solutions is using Bloom filters [21]. This solution applies Bloom filters with keyed hash message authentication codes on q-grams (for a particular string, q-grams gives all possible sub-strings of length q) of identifiers and allows errors in identifiers. Compared to other privacy preserving record linkage methods with encrypted string type identifiers, these methods have lower false negative rates. However, the existing proposed solutions of this category are designed for textual data. On the other hand, privacy preserving record linkage for error-prone numeric data is also very important. For instance, medical records usually contain ample numerical attributes, such as the patient's blood pressure, height, weight, and other test results. In different medical databases, medical data may be stored in different precisions. Then even two very close numbers (eg, 392.1 and 392.11) may cause a totally negative linking result. The consequence of high false negative results may be very harmful, especially when querying a patient's records for emergency treatments.

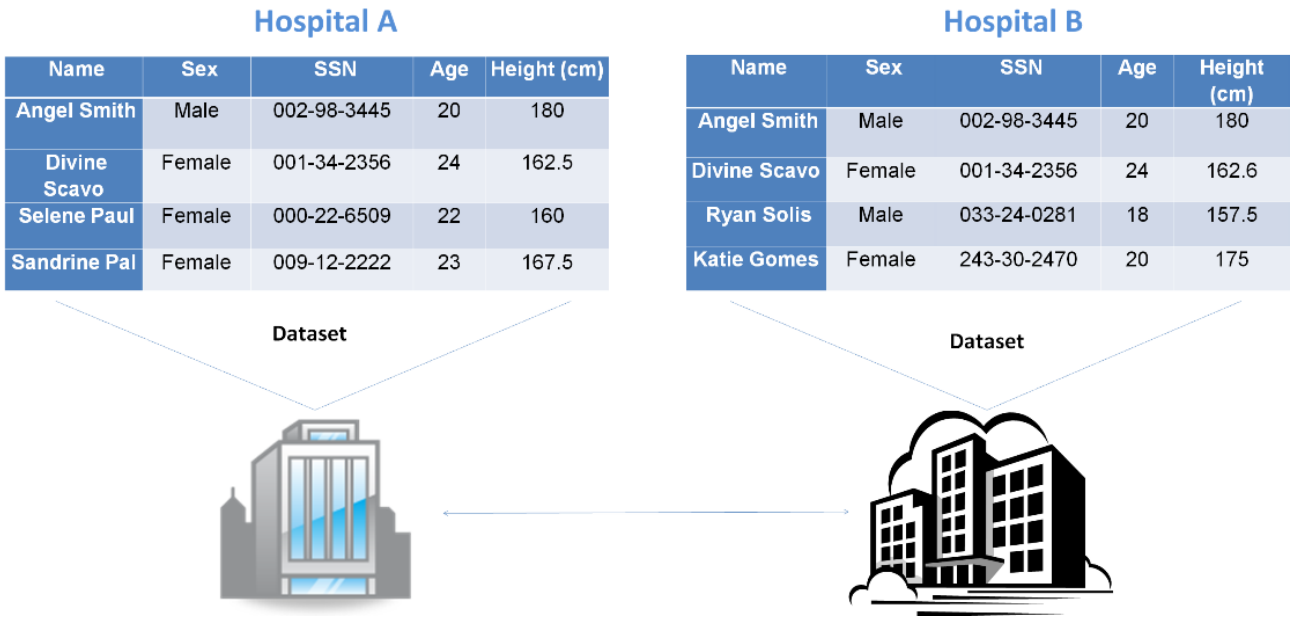
Aim of the Study

In this paper we aim to address the privacy preserving record linkage problem with the presence of error-prone data. In Figure 1 we illustrate an example of real-world cases where privacy preserving record linkage for error-prone data is needed. In this figure, each of the two hospitals holds a database of patients' information of its own. They would like to find out the common patients they share (eg, *Angel Smith* and *Divine Scavo*) in order to perform collaborative research on the shared data. However, due to the requirement of HIPAA, they cannot exchange data in clear texts. Moreover, we notice that for the patient *Divine Scavo*, all attributes are the same at both hospitals except the height (one is 162.5 cm and the other is 162.6 cm). If traditional cryptographic schemes are used, the record belonging to the same patient will be labeled as a mismatch, leading to an error result. In order to avoid the mismatch for the records belonging to the same patient, we need new software to enable privacy preserving record linkage for error-prone data.

In this paper, we designed and developed comprehensive record linkage software for medical organizations, which meets the regulation of HIPAA. Our solution for the privacy preserving record linkage will work not only for error-free data, but also for error-prone numerical data, which is never enabled in existing solutions. The design of our software is open so that the existing textual matching methods (eg, Weber et al) [24] can be easily integrated into the system. Our algorithm used in the software is correct, secure, and efficient. Furthermore, our solution does not need a trusted third party for any of the offered cryptographic schemes. This is important because in many cases such a trusted third party can hardly be found, especially when the health care providers are from different regions or even countries. With a trusted third party added to the software we can use public key cryptographic schemes [25,26]. In particular, we allow the software users to select one of their collaborators who is also using our record linkage software, choose some particular attributes of the database to be matched, and establish the connection with the collaborator after the negotiation. For error-free data, our solution offers two different choices for cryptographic schemes (ie, Secure Hash Algorithm-SHA-1 and SHA-2) [27]. For error-prone data, we propose a new linking algorithm named the Error-Tolerant Linking Algorithm. The Error-Tolerant Linking Algorithm matches the error-prone numerical data and preserves the data privacy for each user too. To achieve this goal, we securely measure the Euclidean distance between two records. If the distance is below a threshold, we say that there is a link between the two records.

It is challenging to compute the distance between two records in a privacy preserving way, such that the other party can learn no information of the original attribute values. To overcome this difficulty, we carefully designed a novel algorithm utilizing the homomorphic property of an efficient cryptographic scheme, the ElGamal, in its extended form. After the linkage process, our solution is also capable of managing the matching records from recent history.

Figure 1. Privacy preserving record linkage problem.



Methods

Design Consideration

This section describes the design consideration of privacy preserving record linkage in general, and the design consideration of record linkage for the error-prone numeric data formally in details. Preserving privacy is a real issue when two or more organizations are willing to share part of their entire data without revealing any sensitive information about any entity to each other. Assume the privacy preserving record linkage takes place between two medical organizations. Each organization holds information about its entities (eg, patients/customers). Along with the different entities, both of these organizations have some common entities too. It is very difficult to get the data of only these common entities from the entire dataset of two organizations while preserving privacy at the same time.

We can explain the overall problem as a real life scenario. For example, suppose privacy preserving record linkage takes place between two hospitals (eg, Hospital A and Hospital B). Figure 2 shows the detailed information/attributes about patients, such as, patient's name, date of birth, address, SSN, sex, etc that each of these hospitals maintains. Hospital A has four patients, *Angel Smith, Divine Scavo, Selene Paul,* and *Sandrine Pal,* and Hospital B has four patients, *Angel Smith, Divine Scavo, Ryan Solis,* and *Katie Gomes*. All the information of patient *Angel Smith* in Hospital A matches with the patient *Angel Smith* in Hospital B. Some of the information (ie, name, address, SSN, age, and sex) of patient *Divine Scavo* in Hospital A matches with patient *Divine Scavo* in Hospital B. Assume that Hospital A is the initiated organization (ie, it takes the initiative of record linkage) and Hospital B is the participating organization as it participates in the record linkage. We use the terms initiated organization and participating organization afterwards in this paper. During this entire procedure it is implicit that Hospital B agrees to share its patients' database with Hospital A without revealing any sensitive information about the patients. Now Hospital A should get *Angel Smith* as matched data, *Divine Scavo* as partially matched data, and *Selene Paul* and *Sandrine Pal* as mismatched data as a result. Note that, here matched data means the data that belongs to Hospital A as well as to Hospital B, mismatched data means data that belongs to Hospital A, but not to Hospital B, and partially matched data means data for an entity of which some of the attributes match at both hospitals' end.

Errors in database data are very usual. Therefore privacy preserving record linkage for error-prone data is necessary too. For error-prone numerical data we can formulate the problem as follows. We assume that for any client, the common part of their records stored in both entities has n attributes. The goal of the linkage is to find out the records held by party B that are within a small distance (very close) to the records held by party A. Formally, the problem for error-prone data (ie, privacy preserving error-tolerant linkage) in this paper can be defined as follows–given two databases D_A ($a_1, a_2, \ldots\ldots, a_m$) and D_B($b_1, b_2, \ldots\ldots, b_m$) with the same attributes. The error-tolerant linkage function takes a tuple <a, D_B, τ> as input, where a is any record in D_A and τ is the distance threshold. It outputs a vector of Boolean numbers, ($r_1, r_2, \ldots, r_m$), where $\forall i$ s.t., $1 \leq i \leq m$ (Figure 3a shows the output vector), in which *Dist* () is the distance function defined for input records (in this paper we use Euclidean distance). Privacy preserving error-tolerant linkage guarantees that computing the error-tolerant linkage function

is secure in the semihonest model [28,29], without a trusted third party.

By being secure in the semihonest model, we mean that the two parties (or any other adversary) cannot efficiently obtain more information than the input and the output of the algorithm. In particular, for our error-tolerant linking algorithm, the two parties will know only the output (r_1, r_2, . . ., r_m) and no information about the values of records (either linked or not-linked) will be revealed.

Figure 2. Data from two hospitals.

Hospital A

Name	Address	Sex	SSN	Age	Body Mass Index	Fasting Blood Sugar
Angel Smith	74 S University Place	Male	002-98-3445	20	24.8	70
Divine Scavo	214 S. West St.	Female	001-34-2254	24	23.3	85
Selene Paul	222 N. Duck St.	Female	000-22-6509	22	22.5	70
Sandrine Pal	200 S. Duncan St.	Female	009-12-2222	23	23.0	72

Hospital B

Name	Address	Sex	SSN	Age	Body Mass Index	Fasting Blood Sugar
Angel Smith	74 S University Place	Male	002-98-3445	20	24.8	70
Divine Scavo	214 S. West St.	Female	001-34-2254	24	23.4	86
Ryan Solis	242 Jade Clover Lane	Male	033-24-0281	18	21.5	71
Katie Gomes	201 Amethyst Lane	Female	243-30-2470	20	24.0	75

Figure 3. Equations (a) Output Vector, (b) Decryption equation of ElGamal scheme, (c) The expanded message of SHA-1, and (d) Proof of correctness of Error-Tolerant Linking Algorithm.

(a) $$r_i = \begin{cases} 1 & Dist(a_i, b_i) \leq \tau \\ 0 & otherwise \end{cases}$$

(b) $$\frac{C_2}{{C_1}^x} = \frac{m.h^r}{g^{x.r}} = \frac{m.h^r}{h^r} = m$$

(c) $$W_i = \begin{cases} M_i & for\ 0 \leq i \leq 15 \\ W_{i-3} \oplus W_{i-8} \oplus W_{i-14} \oplus W_{i-16} & for\ 16 \leq i \leq 79 \end{cases}$$

(d) $$D(C)$$

$$= D\left(\prod_{j=1}^{n} E(g^{a[j]^2})\, E\left(g^{(b_i[j]^2)}\right) E(g^{a[j]})^{-2b_i[j]}\right)$$

$$= D\left(E\left(\prod_{j=1}^{n} g^{(a[j])^2}\, g^{(b_i[j])^2}\right) \prod_{j=1}^{n} E(g^{a[j]})^{-2b_i[j]}\right)$$

$$= D\left(E\left(\prod_{j=1}^{n} g^{(a[j])^2}\, g^{(b_i[j])^2}\right) E\left(\prod_{j=1}^{n} (g^{-2a[j]b_i[j]})\right)\right)$$

$$= D\left(E\left(\prod_{j=1}^{n} g^{(a[j])^2}\, g^{(b_i[j])^2} \prod_{j=1}^{n} (g^{-2a[j]b_i[j]})\right)\right)$$

$$= D\left(E\left(\prod_{j=1}^{n} g^{(a[j])^2}\, g^{(b_i[j])^2}\, g^{-2a[j]b_i[j]}\right)\right)$$

$$= D\left(E\left(g^{\sum_{j=1}^{n} (a[j]-b_i[j])^2}\right)\right)$$

$$= g^{\sum_{j=1}^{n} (a[j]-b_i[j])^2}$$

Privacy Preserving Record Linkage Schemes

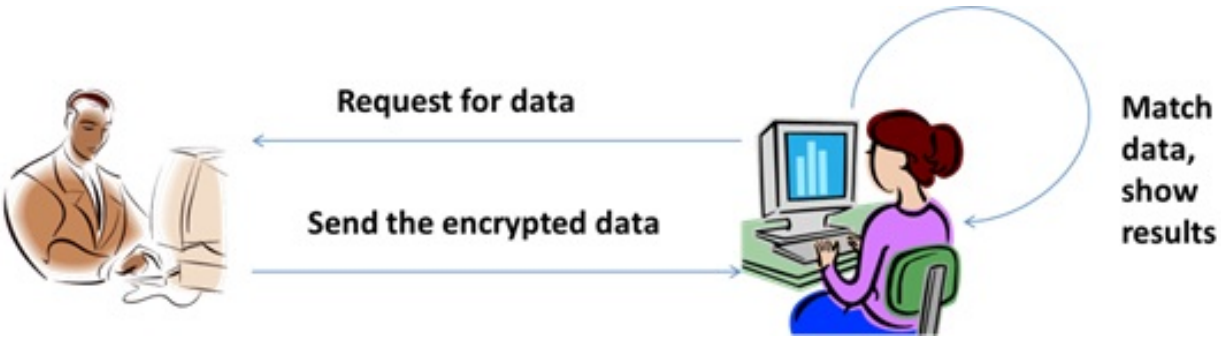
In this subsection, we discuss the overall solution for the design consideration described in the previous section and schemes of the solution in details. The main idea is that if the participating organization sends the entire dataset as encrypted format to the initiated organization, then it is not possible for any other third party, as well as the initiated organization, to know about the real data of the participating organization if the key-pair is unknown. In our proposed solution, to send data confidentially to the other party we have considered three cryptographic schemes: (1) SHA-1, (2) SHA-2, and (3) Error-Tolerant Linking Algorithm. Before discussing the schemes in detail, we categorize the data in two different data categories: (1) the error-free data, and (2) the error-prone data. Among the above-mentioned three cryptographic schemes, the first two (ie, SHA-1 and SHA-2) are the basic cryptographic schemes for privacy preserving error-free data linkage and the Error-Tolerant Linking Algorithm is for the error-prone data.

The overall flow of running the system is as follows. To encrypt the data, the initiated organization chooses a dataset name and the cryptographic scheme, and sends both of them to the

participating organization. If the participating organization holds the same dataset, it starts the privacy preserving data linkage process by sending the data in cipher text format to the initiated organization. Meanwhile, the initiated organization encrypts its own dataset by using the cryptographic scheme. After receiving the data, the initiated organization applies the privacy preserving matching scheme to obtain the results (ie, the matched, mismatched, and partially matched data). Figure 4 shows the diagram of the overall flow of the solution.

We will discuss the Error-Tolerant Linking Algorithm in detail, and the two existing cryptographic schemes briefly in the following two subsections, respectively.

Figure 4. Overall flow of the solution.



Scheme for Error-Prone Data

Our proposed new solution for error-prone data, as above-mentioned, is the Error-Tolerant Linking Algorithm. The Error-Tolerant Linking Algorithm uses the ElGamal [26] scheme as the basic building block. In this subsection, we review the ElGamal scheme first and then will describe the Error-Tolerant Linking Algorithm in detail. The ElGamal is a public key encryption scheme. Let G be a cyclic group of prime order p with generator g. A value $x \in Z_p$ is randomly chosen as the private key. The corresponding public key is (p, g, h), where $h=g^x$. To encrypt the message m, a value $r \in Z^p$ is randomly chosen. Then the cipher text is $E(m) = (C_1, C_2)=(g^r, m.h^r)$. We use $E(m)$ in this paper to denote the cipher text of m encrypted by the ElGamal scheme. The decryption equation of ElGamal scheme is shown in Figure 3b.

Difficulty of computing discrete logarithms over finite fields forms the basis for security in the ElGamal. To decrypt a cipher text, any adversary would have to get the one time random integer. Determining this random integer is infeasible, as it requires computing of discrete logarithms.

The Error-Tolerant Linking Algorithm exploits the homomorphic property of the ElGamal scheme. That is, for two messages m_1 and m_2, it satisfies the following property,

$$E(m_1.m_2)=E(m_1).E(m_2) \quad (1)$$

In addition to linking the data from two different organizations, the Error-Tolerant Linking Algorithm preserves privacy as well. We assume that the attributes of records are preprocessed and converted to integers beforehand. For numerical attributes, this preprocessing is straightforward by normalizing the original values to integers within a certain range. For attributes consisting of strings, we can use a preprocessing method to convert the strings into integers so that the integers can still keep the distance between the records. Then our algorithm can be applied afterwards to complete the records linkage. This algorithm allows the input record with minor deviations less than a small threshold τ. The threshold value is to calculate the distance between the identifiers of two records. In this algorithm, neither entity can learn the records of each other's patients.

Algorithm 1 shows the details of our privacy preserving Error-Tolerant Linking Algorithm. First, party A generates a pair of keys for the ElGamal scheme and sends the public key to party B. For each attribute $a[j]$ in the record, party A computes $g^{a[j]}$ and $g^{((a[j])^2)}$, and sends the cipher texts of these terms to party B. For each record b_i held by party B, party B computes $g^{((bi[j])^2)}$ for each attribute $b_i[j]$, and encrypts them using the public key received from party A. Then party B computes C as shown in line 11 in Algorithm 1. After receiving the product from party B, party A decrypts it using the private key and obtains a decrypted value $D(C)$. If $D(C)$ equals any number in $(g^0, g^1, g^2, \ldots \ldots ., g^\tau)$, then it means that $\sum^n_{k=1}(a[k]-b_i[k])^2 \leq \tau$, and thus we say it is a linking case. Otherwise, we say record b_i does not link to a. The Error-Tolerant Linking Algorithm is correct. We discuss the correctness analysis of the Error-Tolerant Linking Algorithm in the subsection named Correctness Analysis. Figure 5 shows the Error-Tolerant Linking Algorithm–Algorithm 1.

Figure 5. Algorithm 1: Error-tolerant linking algorithm.

INPUT: Party A holds a record$(a[1], a[2], \ldots\ldots, a[n])$. Party B holds a database of m records, and each record in B is of this form: $b_i : (b_i[1], b_i[2], \ldots\ldots, b_i[m])$, where $1 \leq i \leq m$; the distance threshold is τ.

OUTPUT: Linking result for each record held by B: $[r_1, r_2, \cdots, r_m]$, where $r_i = 1$ or $r_i = 0$, $1 \leq i \leq m$.

1. Party A generates a pair of public key (p, g, h) and private key x for ElGamal scheme and sends (p, g, h) to B.
2. for j= 1 through n do
3. A computes $g^{a[j]}$ and encrypts $g^{a[j]}$, obtaining $E(g^{a[j]})$.
4. end for
5. A computes $E(g^{(a[1])^2})$, $E(g^{(a[2])^2})$, ..., $E(g^{(a[n])^2})$ and sends them together with $E(g^{a[1]})$, $E(g^{a[2]})$, ..., $E(g^{a[n]})$ to B.
6. for i = 1 through n do
7. $r_i = 0$
8. for j = 1 through m do
9. B computes $g^{((b_i[j])^2)}$ and encrypts $g^{((b_i[j])^2)}$, obtaining E $(g^{((b_i[j])^2)})$.
10. end for
11. B computes $\prod_{j=1}^{n} E(g^{a[j]^2})\, E(g^{(b_i[j]^2)}) E(g^{a[j]})^{-2b_i[j]}$ and sends it to A.
12. A decrypts C using private key x, and obtains $D(C) = g^{\sum_{j=1}^{n}(a[j]-b_i[j])^2}$.
13. for k = 1 through τ
14. if $D(C) = g^k$ then
15. Output result $r_i = 1$.
16. Break
17. end if
18. end for
19. end for

Schemes for Error-Free Data

For the regular error-free data, we have considered two existing basic cryptographic schemes: (1) SHA-1, and (2) SHA-2. The SHA-1 [27] and SHA-2 [27] come under the hash algorithm family. In this subsection, we will review the SHA briefly. The SHA is based on the design principle of the Message Digest Algorithm 4 (MD4) [29]. Both of these algorithms are iterative and one-way hash functions. The SHA-1 and SHA-2 consist of two major steps: (1) preprocessing, and (2) hash computation. In the preprocessing step, every input message is padded and then split into fixed size message blocks, and this step also initializes the working variables to be used in hash computation. The hash computation consists of an 80-step compression function that iteratively generates hash values h_i (ie, the i^{th} hash value). The 80-step compression function is applied to each of the message blocks. Generally, two types of inputs are considered here: (1) chaining input, and (2) message. If the message is m and chaining input is h_i, then the compression function is $g(m, h_i)$ at the i^{th} stage. The chaining input $h_{(i+1)}$ at the $(i+1)^{th}$ stage is calculated by $h_i+g(m, h_i)$. The value of the compression function at the last stage is the hash value of the message. The SHA-1 and SHA-2 differ in terms of the message size, block size, word size, and message digest size as given in Table 1.

In the SHA-1, five working variables are used: (1) a, (2) b, (3) c, (4) d, and (5) e. The message is represented by 16 32-bit words, denoted by M_i. The message is then expanded to 80 32-bit words W_i. The expanded message $W_{(i)}$ is shown in Figure 3c.

After that it initializes the working variables and computes the 80-step compression function and intermediate hash values. If there are n message blocks (ie, M_1, M_2, … . ., M_n), then the entire procedure is repeated for n number of times. The resulting 160-bit message digest of the message M is,

$$H_0^{(N)} || H_1^{(N)} || H_2^{(N)} || H_3^{(N)} || H_4^{(N)} \quad (2)$$

Here, $H_j^{(i)}$ means the j^{th} word of i^{th} hash value.

The procedure of the SHA-2 is similar to the SHA-1. It first pads the message and divides it into 64-bit message blocks. The number of working variables here are eight (ie, a, b, c, d, e, f, g, and h). After initializing the working variables and computing the 80-step compression function and intermediate hash values, it generates 512-bit message digest of the message M. The final message digest of M is,

$$H_0^{(N)} || H_1^{(N)} || H_2^{(N)} || H_3^{(N)} || H_4^{(N)} || H_5^{(N)} || H_6^{(N)} || H_7^{(N)} \quad (3)$$

The SHA-1 and SHA-2 are considered here since it is easy to compute the hash value of any given message and they are the one-way hash functions (ie, they have one-way, second preimage resistant, and collision resistant properties). The SHA-1 and SHA-2 produce 160-bit and 512-bit hash values, respectively, for any given message. Therefore, for any given message, there are 2^{160} and 2^{512} possible hash values. It is very difficult to identify the actual message from this vast range of hash values [30]. Here in our system, after applying the SHA-1 and SHA-2 on the data, we get the message digest/encrypted data from both of the parties and apply data matching techniques on those encrypted data.

Table 1. The SHA-1 and SHA-2 properties.

Secure hash algorithm name	Message size (bits)	Block size (bits)	Word size (bits)	Message digest size (bits)
SHA-1	$<2^{64}$	512	32	160
SHA-2	$<2^{128}$	1024	64	512

System Analysis

We analyze our schemes, especially the Error-Tolerant Linking Algorithm, in terms of correctness, privacy, and complexity.

Correctness Analysis

For the proof of correctness, if the two parties follow Algorithm 1, they will jointly compute the correct Euclidean distance without each party knowing the record from the other party. The homomorphic property of the ElGamal scheme helps to prove line 12 in Algorithm 1. (Figure 3d shows the proof of correctness of Error-Tolerant Linking Algorithm.

If $(C)=g^k$, where $k \leq \tau$, it means that $\sum^{n}_{k=1}(a[k]\text{-}b_i[k])^2 \leq \tau$. Then we can say that record a is within the distance of τ, from record b_i, and the result of error-tolerant linking is positive.

Privacy Analysis

In this subsection, we explain why the Error-Tolerant Linking Algorithm is secure (ie, privacy preserving) in the semihonest model. Being secure in the semihonest model means neither of the two parties can learn more than the output of the algorithm from the information received during the algorithm. In Algorithm 1, the only message received by B is the cipher texts $E(g^{((a[1])^2)})$, $E(g^{((a[2])^2)})$,…, $E(g^{((a[n])^2)})$, and $E(g^{a[1]})$, $E(g^{a[2]})$,…, $E(g^{a[n]})$. Since the ElGamal scheme is semantically secure under the decisional Diffie-Hellman assumption [31], party B cannot learn anything about $g^{((a[1])^2)}$, $g^{((a[2])^2)}$,…, $g^{((a[n])^2)}$, $g^{a[1]}$, $g^{a[2]}$,…, $g^{a[n]}$ but the cipher texts. For party A, the message it receives from party B is C. From the semantic security of the ElGamal scheme, party A cannot learn the clear texts from party B but the $D(C)$. Here we note that $D(C)$ and $\sum^{n}_{k=1}(a[k]\text{-}b_i[k])^2$ can be derived from the output of the algorithm by trying different numbers in a small range of τ. Therefore, we say that party A knowing $D(C)$ and $\sum^{n}_{k=1}(a[k]\text{-}b_i[k])^2$ does not violate the security requirement and party A can send these values to party B if needed.

Complexity Analysis

We analyze the computation cost of our algorithm on party A and party B respectively. The computation cost of party A includes computing $2n$ exponentiations, $2n$ encryptions, and 1 decryption. Suppose that each exponentiation takes time T_e. Then the total computation cost of party A is $2n(T_e+T_E)+T_D$, where T_E is the time to perform one ElGamal encryption and T_D is the time to perform one decryption. Values of g^k can be computed beforehand and saved in a table for reference at line 14, Algorithm 1. Party B needs to compute $2nm$ exponentiations, nm encryptions, and $2nm$ divisions/multiplications. So the computation cost for party B is $nm(2T_e+T_E+2T_m)$, where T_m is time for a multiplication, and the definitions of T_e, T_E are as above. When there are k records in party A to be linked with the records held by party B, the total computation time for party A is $nm(2T_e+T_E)+2nmk \times T_m$. Note that in the computation cost, $nm(2T_e+T_E)$ is not multiplied by k, because as long as party A does not change their public key, the cipher texts of party B's records do not change, and thus they only need to be computed once.

Results

System Implementation

This subsection explains the system implementation we have taken into account for the problem described before. We implement our system using the Eclipse Integrated Development Environment. We have used the programming language Java. The entire system is divided into three modules: (1) Connection Management Module, (2) Data Matching Module, and (3) Matching Record Management Module. Among these three modules, the main module is Data Matching Module. The solution of the privacy preserving record linkage (ie, Data Matching Module) works for both the error-free data and error-prone numerical data. The Matching Record Management Module shows the result/records from the recent past data matching attempts, and the Connection Management Module takes care of creating a connection with the collaborator. We will discuss these three modules in detail in the following subsections. For the snapshot of selecting a function/module in our system, see Multimedia Appendix 1.

Connection Management Module

Each party/organization keeps a list of available and reachable collaborators. To create a connection with another party/collaborator, each party needs to select that particular collaborator from the collaborator list. For the snapshot of how a user selects a collaborator in our system see Multimedia Appendix 2. To initialize the connection, each and every organization keeps some initial information about the other collaborators beforehand. This information contains the Internet Protocol (IP) address, port number, public-key, private-key pair, etc. Party A first selects party B from the available participating collaborator list. Party A uses the corresponding IP address and port number of party B for creating a connection. We follow the client-server architecture to implement our system. The communication between two parties is realized by socket application program interface (API). Party B (server) creates a socket to listen to requests from party A (client). Party B can handle more than one client at a time. In that case, party B creates a separate socket for each of the requesting clients using multi-threading. To be precise, a user can work as both a client and a server at the same time. A user can turn on the server and continue working as a client using the data matching procedure.

Data Matching Module

The Data Matching Module is the main module of our system. To explain this module, we consider there are two parties: (1) party A, and (2) party B. Party A initiates the matching procedure and party B takes part in this matching procedure. Figure 6 shows the entire workflow of the system including party A and party B.

As shown in Figure 6, once a connection is created between party A and party B, data transfer between the two parties and matching can take place. Party A first selects the record set for matching data. When party A selects the dataset name, then the corresponding attributes' list becomes available. Party A selects the attributes' names and sends the dataset name along with the attributes to party B. Party B searches the requested record set in its set of record sets. If party B has the record set, it sends the acknowledgement to party A. In response to this acknowledgement, party A sends the user selected cryptographic scheme name to party B and encrypts its own selected record set. Party B encrypts the requested dataset with the requested cryptographic scheme. Figure 7 shows the snapshot of how a user selects a cryptographic scheme. For encryption purposes, the Java cryptography library is used. To encrypt data, we have considered three cryptographic schemes: (1) SHA-1, (2) SHA-2, and (3) Error-Tolerant Linking Algorithms. The first two schemes, SHA-1 and SHA-2, do not require any key pair, whereas the Error-Tolerant Linking Algorithms needs a key pair for encryption. As of now, we have considered that the organization and its collaborator will know the key pair beforehand. The first two schemes, SHA-1 and SHA-2, work in the same way. After encrypting the record set, party B sends the encrypted data to party A. After receiving the encrypted data from party B, party A applies the data matching technique on these two encrypted record sets.

The Error-Tolerant Linking Algorithm works in a little bit different way than the other two cryptographic schemes once the dataset name and attributes have been selected. Suppose party A has one medical record to be linked with the records held by party B. (The flow can be easily extended to the cases that party A has multiple medical records to be linked.) Party A sends the encrypted messages generated by their data record to be linked to party B. Then party B handles the encrypted messages as described in previous sections (ie, encrypting their own data record and multiplying their inverse with party A's message) and sends the multiplied encrypted message back to party A. Party A decrypts the message and outputs the linking result. Party B moves to the next record and repeats the linking procedure. Party A does not need to encrypt their record again, but only needs to decrypt the messages sent from party B and output the linking result.

This repeating process carries on until party B has gone through all their records. Then party A and party B close the connection with each other and the privacy preserving linkage is completed. Figure 8 shows the flow of the Error-Tolerant Linking Algorithm after selecting the dataset name and attributes. Once the entire data matching procedure is completed, the client/initiating organization closes the connection with the server/participating organization automatically. As a result of the entire data matching procedure, party A gets the matched, mismatched, and partially matched result with party B. Figure 9 shows the snapshot of the matching result of our system.

Figure 6. The workflow of the system.

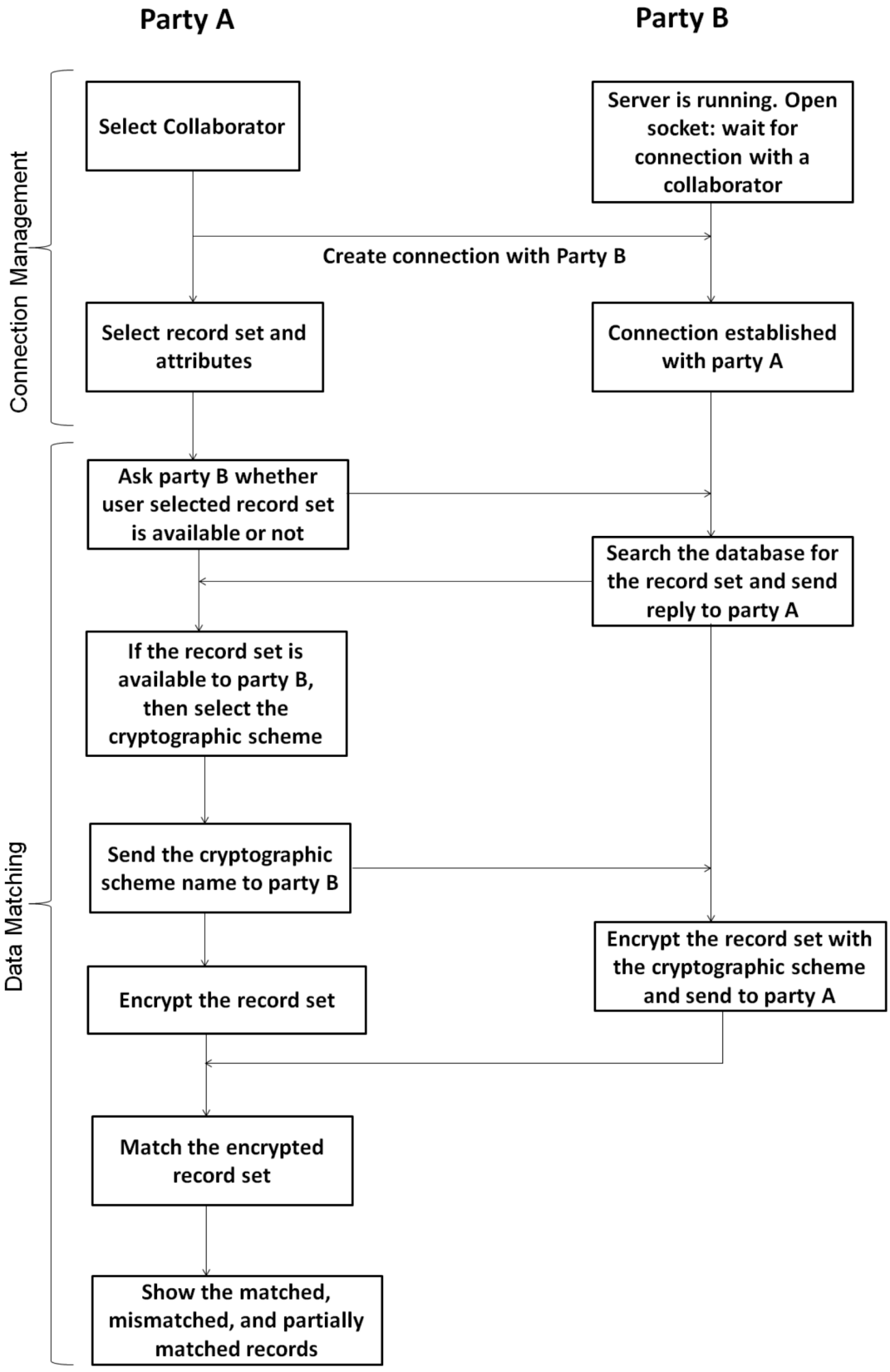


Figure 7. Snapshot of selecting a cryptographic scheme.

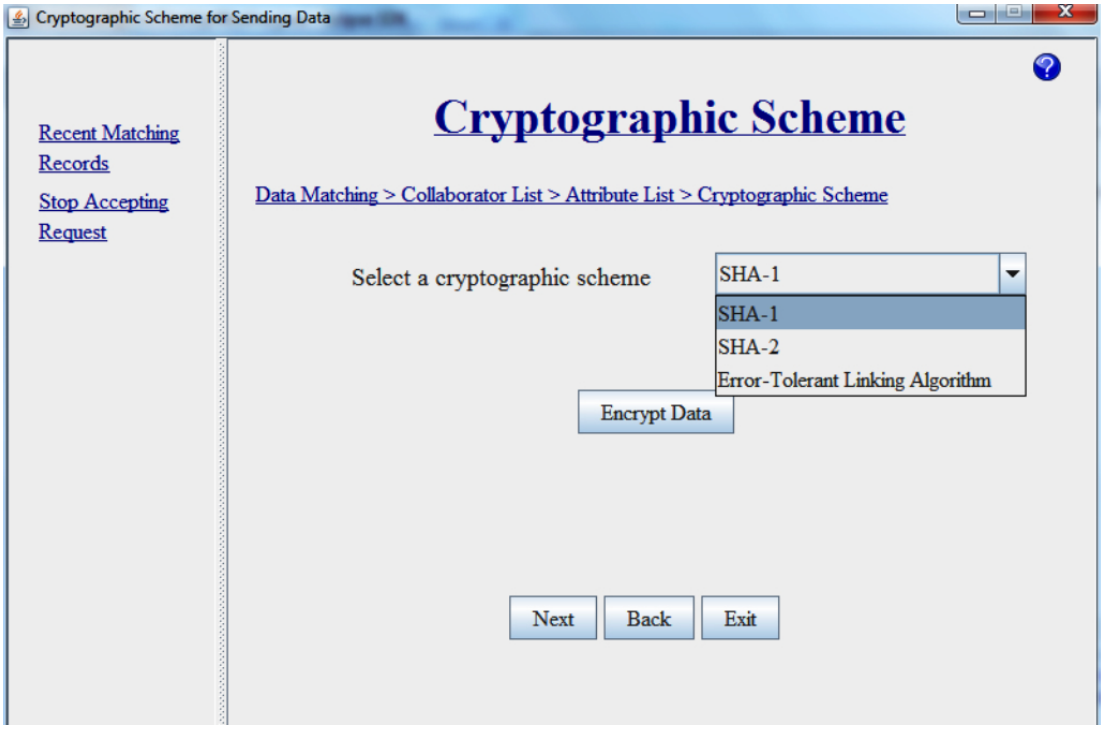


Figure 8. Flow of the Error-Tolerant Linking Algorithm between two parties when they are already connected and the dataset name and attributes are known to party B.

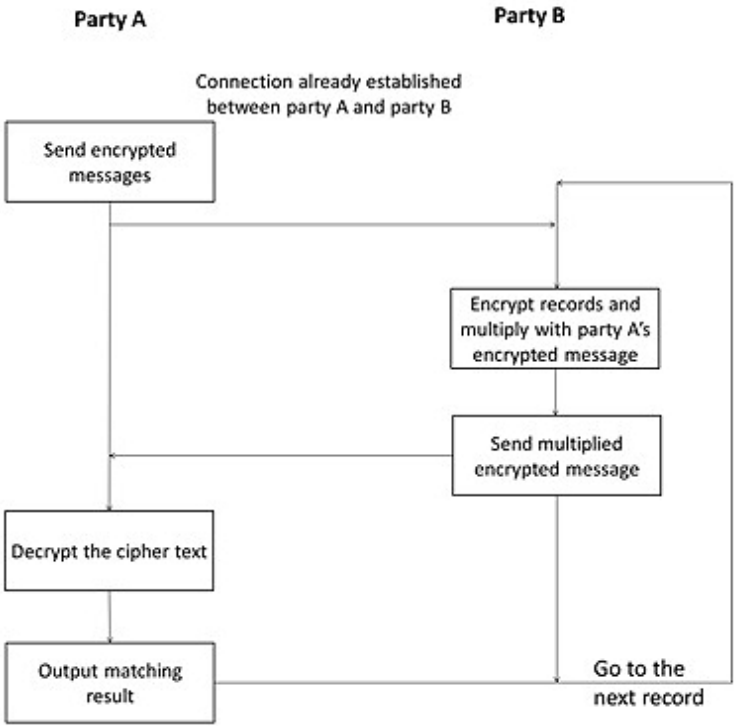
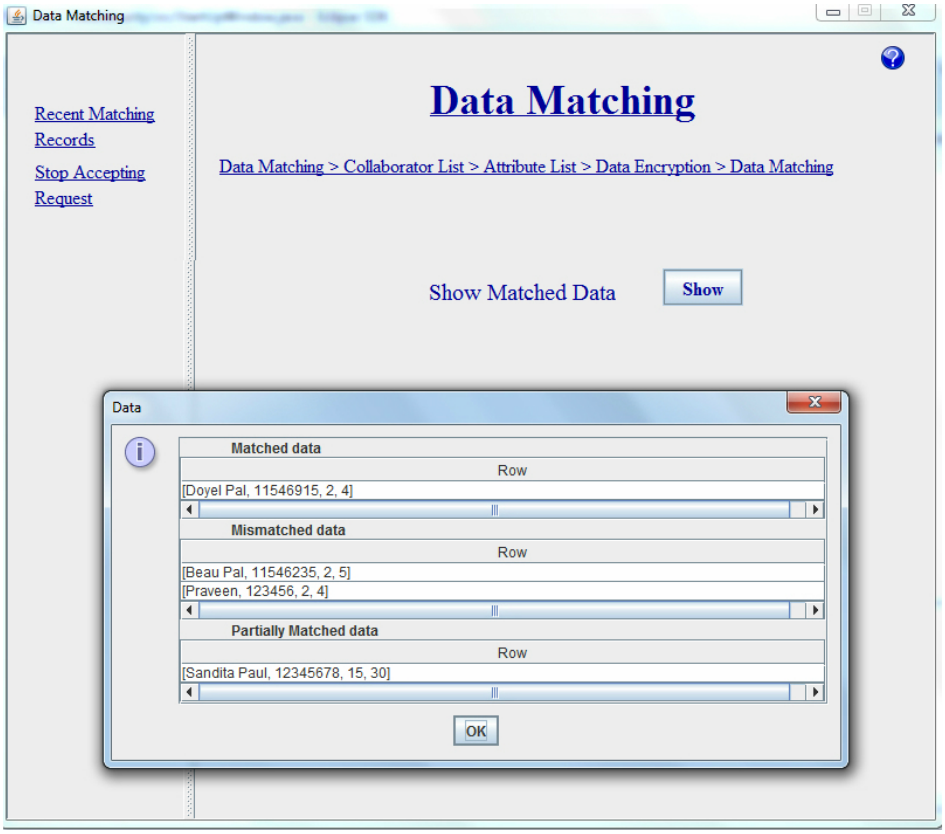


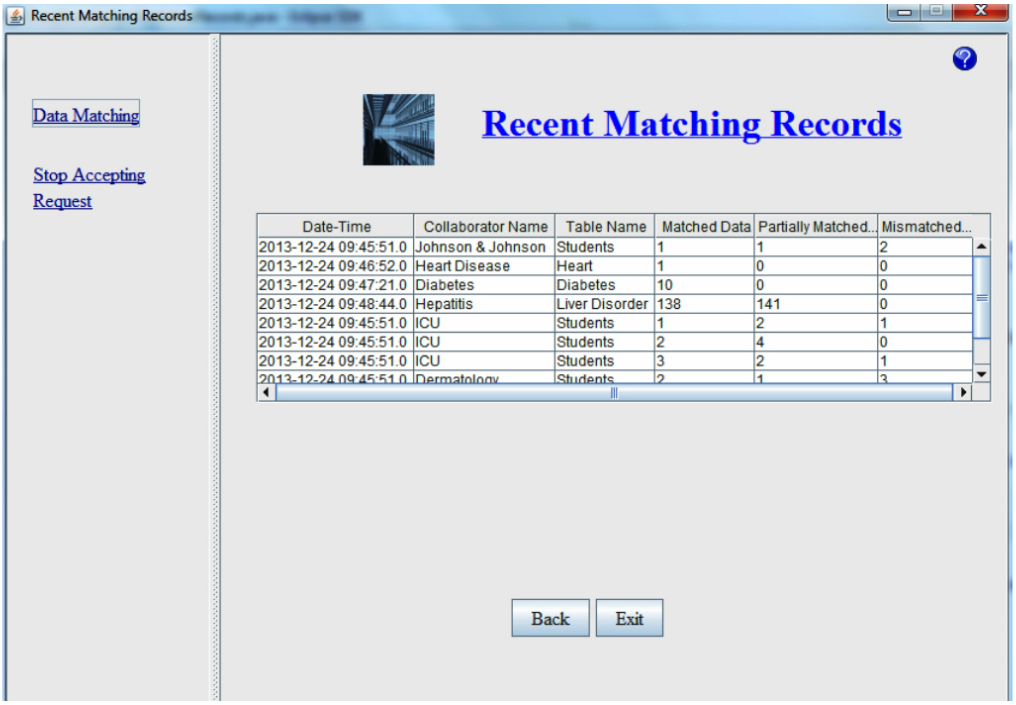
Figure 9. Snapshot showing the matching result in our system.



Matching Record Management Module

The Matching Record Management Module shows the brief description of the matching result from the recent past data matching attempts. It shows the date-time of when the matching took place, name of the participating organization/collaborator, name of the record set, and the number of matched, mismatched, and partially matched data for each and every data matching attempt in table format. Figure 10 shows the snapshot of the Matching Record Management Module of our system.

Figure 10. Matching Record Management Module.



Experiment Setup

We ran our system on computers with a 3.33 GHz Intel Core i5 processor with 4 GB RAM and a 64 bit operating system. Both for party A and party B, we have constructed window applications using Java. The Internet connects the applications on different computers. The communications between party A and party B are realized by using socket API. Before running the system, each client needs to know the IP address and port number of the server. If a party/server changes their IP address, then they should inform the other parties/clients. As of now, we have considered that each party maintains an IP address, and port list of other parties.

We use two real-world medical datasets, the Pima Indians Diabetes Data Set and the Heart Disease Data Set [32] to implement our system. To handle these real-world datasets, MySQL (an open-source database system) is used. Java Database Connectivity helps to connect the application front end and the database end. It is used to access data directly from the database and to show them to the user.

Experimental Results

We test the scalability of our system in terms of time efficiency. For each cryptographic scheme in this system, we vary the number of records and the number of attributes for each record, and then measure the computation time of our system.

To test the efficiency of our system, we consider two real-world datasets, the Pima Indians Diabetes Data Set and the Heart Disease Data Set [32]. In the Pima Indians Diabetes Data Set, we use at most eight attributes: (1) number of times pregnant, (2) plasma glucose concentration; a 2 hours in an oral glucose tolerance test, (3) diastolic blood pressure (mm Hg), (4) triceps skin fold thickness (mm), (5) 2-hours serum insulin (mm U/ml), (6) body mass index (weight in kg/height in m^2), (7) diabetes pedigree function, and (8) age (years). Similarly for the Heart Disease Data Set, we use eight attributes for each record: (1) age of the patient, (2) sex, (3) chest pain type, (4) resting blood pressure, (5) serum cholesterol in mg/dl, (6) fasting blood sugar, (7) resting electrocardiographic results, and (8) maximum heart rate achieved. For each encryption scheme, except the Error-Tolerant Linking Algorithm, we vary only the number of attributes to four, six, and eight and use 100 patients' records. For the Error-Tolerant Linking Algorithm, we vary the number of attributes as well as the number of patients' records.

For the SHA-1 and SHA-2, we use 100 patients' records from both the Pima Indians Diabetes Data Set and Heart Disease Data Set. For each record, we vary the number of attributes to four, six, and eight respectively. Figures 11 and 12 show the computation times of our system using the SHA-1 and SHA-2. For both of these above-mentioned existing algorithms, the computation time increases as we increase the number of attributes. The computation time grows almost linearly as we increase the number of attributes. Moreover, in every case the computation time does not even go beyond 0.1 second.

Figures 13 and 14 show the computation time for the Error-Tolerant Linking Algorithm. We implement the Error-Tolerant Linking Algorithm with both constant and varying numbers of attributes. The values of all the attributes are preprocessed and converted to integers.

Figure 13 shows the computation time of our system when party A conducts the privacy preserving linking on the Pima Indians Diabetes Data Set. Party B holds the variable number of records varying from 100, 200, and 300, while keeping number of attributes constant as four. The computation time increases linearly as the size of data to be linked grows. Figure 14 presents the computation time for the Heart Disease Data Set with varying numbers of records and varying numbers of attributes. For this data set too, the computation time increases as the number of attributes and number of records grow. In both cases, when the number of records or number of attributes increases, the computation time increases almost linearly. In addition to that, for this algorithm too, the computation times never go beyond 0.1 second.

Figure 11. Computation time for SHA-1 for 100 records with varying number of attributes to four, six, and eight.

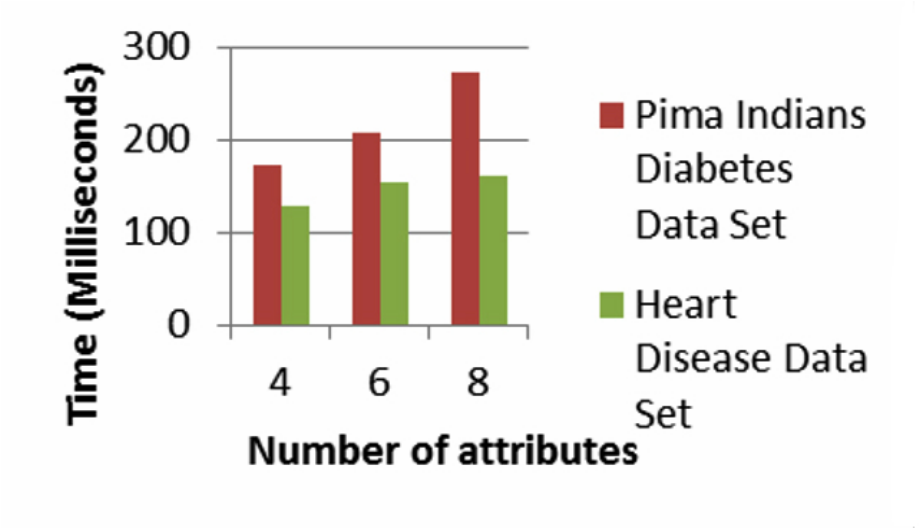


Figure 12. Computation time for SHA-2 for 100 records with varying number of attributes to four, six, and eight.

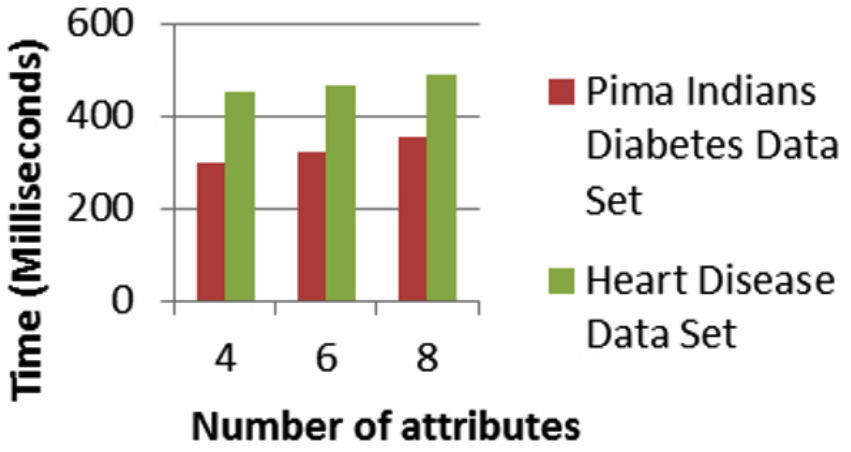


Figure 13. Computation time for the Error-Tolerant Linking Algorithm with varying patients' records from Pima Indians Diabetes Data Set where each record has 4 attributes.

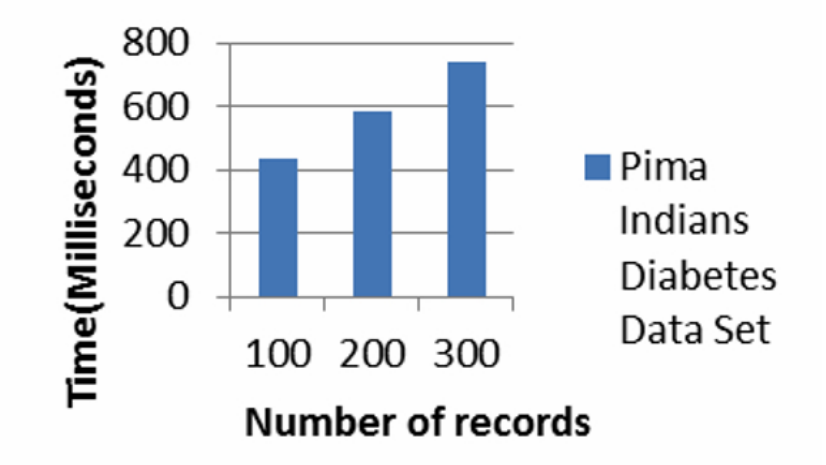
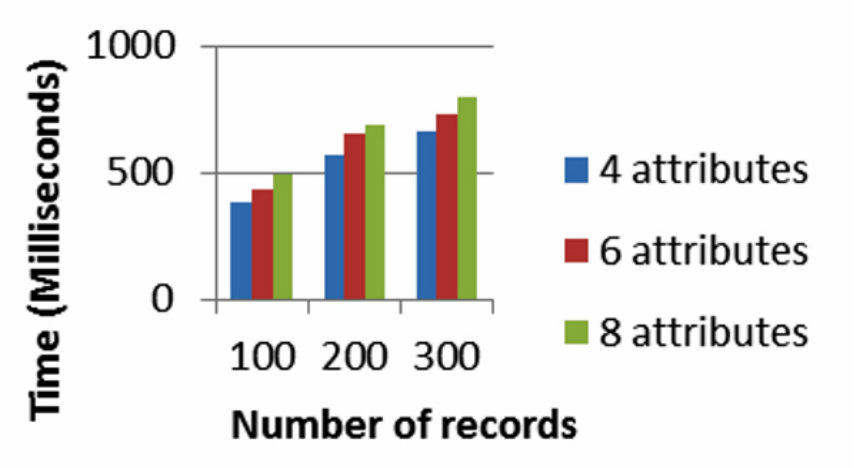


Figure 14. Computation time for the Error-Tolerant Linking Algorithm with varying patients' records and varying attributes from Heart Disease Data Set.



Discussion

Principal Findings

To enhance the health care quality and public health surveillance, privacy preserving medical record linkage among different medical service providers is very important. As the real-world medical record may well be error-prone, the goal of our study was to design and develop a software system that helps medical record linkage for both error-free data and error-prone data, and preserves privacy too. We have successfully designed a comprehensive system to achieve this goal. Moreover, our software meets the regulation of HIPAA and does not require a trusted third party. Our software preserves privacy since no party can get to know about another's database. As the existing works on error-prone data are limited to textual data, we propose a novel algorithm named the Error-Tolerant Linking Algorithm, which works on error-prone numeric data. We offer two cryptographic schemes, the SHA-1 and SHA-2 for error-free data. We designed our software open and each cryptographic scheme is independent to each other so that any existing work/cryptographic scheme for error-prone textual data can be integrated later. We tested our system on real-world datasets and got the expected result each time for each of the offered cryptographic schemes. Besides that, our system is efficient for real-world datasets and the computation time for each attempt has never gone beyond 0.1 second.

Limitations

The one limitation of our proposed system is that for error-prone data our system is limited to only numeric data. Considering this fact, we designed our software in such a way that any existing solution for error-prone textual data can be easily integrated into our system. This makes our software flexible and open to integrate any existing record linkage scheme for error-prone textual data.

Conclusions

In this paper, we propose a solution for privacy preserving record linkage for error-free data as well as for error-prone data. For error-free data, we offer two existing cryptographic schemes: (1) SHA-1, and (2) SHA-2. A new algorithm is proposed for error-prone numeric data. We implement our system fully and tested it on two real-world data sets. We have shown that our system is secure, correct, and efficient and does not require a trusted third party. The experimental results demonstrate the efficiency of our system.

Acknowledgments

This work is partly supported by the Oklahoma Center for the Advancement of Science and Technology, Health Research project HR13-035. SZ was supported in part by the National Science Foundation of China (NSFC)-61321491 and NSFC-61300235. Part of the work was done while both TC and SZ were at the University at Buffalo and supported by the National Science Foundation grant 0845149.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Snapshot of selecting a function in our system.

[JPG File, 11KB - medinform_v2i1e1_app1.jpg]

Multimedia Appendix 2

Snapshot of selecting a collaborator.

[JPG File, 14KB - medinform_v2i1e1_app2.jpg]

References

1. Vest JR. Health information exchange and health care utilization. J Med Syst 2009 Jun;33(3):223-231. [Medline: 19408456]
2. Shapiro JS, Kannry J, Lipton M, Goldberg E, Conocenti P, Stuard S, et al. Approaches to patient health information exchange and their impact on emergency medicine. Ann Emerg Med 2006 Oct;48(4):426-432. [doi: 10.1016/j.annemergmed.2006.03.032] [Medline: 16997679]
3. HealthIT.gov. Nationwide health information network URL: http://www.healthit.gov/policy-researchers-implementers/nationwide-health-information-network-nwhin [accessed 2013-11-05] [WebCite Cache ID 6KwBjGB9j]
4. US Department of Health and Human Services. The privacy rule URL: http://www.hhs.gov/ocr/hipaa/finalreg.html [accessed 2013-11-05] [WebCite Cache ID 6KwBXLREA]
5. Bellare M, Canetti R, Krawczyk H. UCSD CSE-Computer science and engineering. 1996. Keying hash functions for message authentication URL: http://cseweb.ucsd.edu/~mihir/papers/kmd5.pdf [accessed 2013-12-25] [WebCite Cache ID 6MBCiPwD5]

6. Chen T, Zhong S. Proceedings of the AMIA. 2010. An efficient privacy preserving method for matching patient data across different providers URL: http://knowledge.amia.org/amia-55142-a2010-1.115988/annual2010-1.116530/f-001-1.116531/a-562-1.116708?qr=1 [accessed 2013-12-27] [WebCite Cache ID 6MCtI1z2D]
7. Quantin C, Bouzelat H, Dusserre L. A computerized record hash coding and linkage procedure to warrant epidemiological follow-up data security. Stud Health Technol Inform 1997;43 Pt A:339-342. [Medline: 10179568]
8. Quantin C, Bouzelat H, Allaert FA, Benhamiche AM, Faivre J, Dusserre L. How to ensure data security of an epidemiological follow-up: Quality assessment of an anonymous record linkage procedure. Int J Med Inform 1998 Mar;49(1):117-122. [Medline: 9723810]
9. Alhaqbani B, Fidge C. IEEE Explore Digital Library. Singapore: e-health Networking, Applications, and Services. HealthCom; 2008. Privacy-preserving electronic health record linkage using pseudonym identifiers URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4600120 [WebCite Cache ID 6MBFTZkSy]
10. Churches T, Christen P. Some methods for blindfolded record linkage. BMC Med Inform Decis Mak 2004 Jun 28;4:9 [FREE Full text] [doi: 10.1186/1472-6947-4-9] [Medline: 15222890]
11. Lindell Y, Pinkas B. ICT. Berlin; 2000. Privacy preserving data mining URL: http://www.ict.griffith.edu.au/~vlad/teaching/kdd.d/readings.d/lindell00privacy.pdf [accessed 2013-12-26] [WebCite Cache ID 6MBdaaRS5]
12. Agrawal R, Srikant R. ACM Sigmod Record 29.2. 2000. Privacy-preserving data mining URL: http://www.utdallas.edu/~muratk/courses/privacy08f_files/agrawal00privacypreserving.pdf [accessed 2013-12-27] [WebCite Cache ID 6MBHah8MY]
13. Goldreich O. Volume 2-basic applications. In: Foundations of cryptography. New York: Cambridge University Press; 2009.
14. Scannapieco M. Privacy preserving schema and data matching. 2007 Presented at: Proceedings of the ACM SIGMOD International Conference on Management of Data; ACM; 2007; New York, NY, USA URL: https://www.cs.purdue.edu/homes/ake/pub/main.pdf
15. Hjaltason GR, Samet H. Properties of embedding methods for similarity searching in metric spaces. IEEE Trans. Pattern Anal. Machine Intell 2003 May;25(5):530-549. [doi: 10.1109/TPAMI.2003.1195989]
16. Inan A, Kantarcioglu M, Bertino E, Scannapieco M. Data Engineering: ICDE. 2008. A hybrid approach to private record linkage URL: http://dl.acm.org/citation.cfm?id=1547255 [accessed 2013-12-27] [WebCite Cache ID 6MCuQdrab]
17. Zhang W, Wang X, Lu B, Kim TH. Secure encapsulation and publication of biological services in the cloud computing environment. Biomed Res Int 2013;2013:170580 [FREE Full text] [doi: 10.1155/2013/170580] [Medline: 24078906]
18. Gkoulalas-Divanis A, Loukides G. SDM. 2012 Apr. Privacy-preserving medical data sharing URL: http://www.siam.org/meetings/sdm12/gkoulas_loukides.pdf [accessed 2013-12-27] [WebCite Cache ID 6MBKN5GLp]
19. Kum HC, Krishnamurthy A, Machanavajjhala A, Reiter MK, Ahalt S. Privacy preserving interactive record linkage (PPIRL). J Am Med Inform Assoc 2013 Nov 7. [doi: 10.1136/amiajnl-2013-002165] [Medline: 24201028]
20. Mohammed N, Jiang X, Chen R, Fung BC, Ohno-Machado L. Privacy-preserving heterogeneous health data sharing. J Am Med Inform Assoc 2013 May 1;20(3):462-469. [doi: 10.1136/amiajnl-2012-001027] [Medline: 23242630]
21. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Mak 2009;9:41 [FREE Full text] [doi: 10.1186/1472-6947-9-41] [Medline: 19706187]
22. Bachteler T, Schnell R, Reiher J. Willkommen an der Universität Duisburg-Essen. 2010. An empirical comparison of approaches to approximate string matching in private record linkage URL: https://www.uni-due.de/~hq0215/documents/Draft_Ottawa_Comparison.pdf [accessed 2013-12-27] [WebCite Cache ID 6MBGp6Pww]
23. Wang J, Donnan PT. Journal of Applied Statistics 29.6. 2002. Adjusting for missing record linkage in outcome studies URL: http://www.tandfonline.com/doi/abs/10.1080/02664760220136186?journalCode=cjas20#preview [WebCite Cache ID 6MBHCEipf]
24. Weber SC, Lowe H, Das A, Ferris T. A simple heuristic for blindfolded record linkage. J Am Med Inform Assoc 2012 Jun;19(e1):e157-e161 [FREE Full text] [doi: 10.1136/amiajnl-2011-000329] [Medline: 22298567]
25. Shamir A, Adleman L, Rivest RL. Commun. ACM. 1978. A method for obtaining digital signatures and public-key crypto systems URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.40.5588 [accessed 2013-12-27] [WebCite Cache ID 6MB91EPdr]
26. Elgamal T. A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Trans. Inform. Theory 1985 Jul;31(4):469-472. [doi: 10.1109/TIT.1985.1057074]
27. National Institute of Standards and Technology (NIST). Computer Security Division Computer Security Resource Center. Federal Information Processing Standards (FIPS) Publications URL: http://csrc.nist.gov/publications/fips/fips180-2/fips180-2withchangenotice.pdf [accessed 2013-11-05] [WebCite Cache ID 6KwC11Gr1]
28. Boneh D, Goh EJ, Nissim K. Evaluating 2-DNF formulas on ciphertexts. In: Theory of cryptography. Berlin: Springer; 2005:325-341.
29. Rivest R. Internet Engineering Task Force. 1992. The MD4 message-digest algorithm URL: http://www.ietf.org/rfc/rfc1186.txt [accessed 2013-12-25] [WebCite Cache ID 6MBJv4km9]
30. Burr WE. Cryptographic hash standards: Where do we go from here? IEEE Secur. Privacy Mag 2006 Mar;4(2):88-91. [doi: 10.1109/MSP.2006.37]
31. Boneh D. Stanford. 1998. The decision Diffie-Hellman problems URL: http://crypto.stanford.edu/~dabo/abstracts/DDH.html [accessed 2013-12-26] [WebCite Cache ID 6MBIc1QoS]

32. UCI Machine Learning Repository. URL: http://archive.ics.uci.edu/ml/datasets.html [accessed 2013-11-06] [WebCite Cache ID 6KwBqigQa]

Abbreviations

API: application program interface
HIPAA: Health Insurance Portability and Accountability Act
IP: Internet protocol
MD4: Message Digest Algorithm 4
NSFC: National Science Foundation of China
SHA: Secure Hash Algorithm
SMC: secure multi-party computation
SSN: Social Security Number

Edited by G Eysenbach; submitted 08.11.13; peer-reviewed by Z Hao, W Zhang; comments to author 03.12.13; revised version received 13.12.13; accepted 14.12.13; published 20.01.14.

Please cite as:
Pal D, Chen T, Zhong S, Khethavath P
Designing an Algorithm to Preserve Privacy for Medical Record Linkage With Error-Prone Data
JMIR Med Inform 2014;2(1):e2
URL: http://medinform.jmir.org/2014/1/e2/
doi:10.2196/medinform.3090
PMID:25600786

Original Paper

Next Generation Phenotyping Using the Unified Medical Language System

Tomasz Adamusiak[1], MD, PhD; Naoki Shimoyama[1]; Mary Shimoyama[1,2], PhD

[1]Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, WI, United States

[2]Department of Surgery, Medical College of Wisconsin, Milwaukee, WI, United States

Corresponding Author:
Tomasz Adamusiak, MD, PhD
Human and Molecular Genetics Center
Medical College of Wisconsin
8701 Watertown Plank Road
Milwaukee, WI, 53226
United States
Phone: 1 4149554995
Fax: 1 4144566516
Email: tomasz@mcw.edu

Abstract

Background: Structured information within patient medical records represents a largely untapped treasure trove of research data. In the United States, privacy issues notwithstanding, this has recently become more accessible thanks to the increasing adoption of electronic health records (EHR) and health care data standards fueled by the Meaningful Use legislation. The other side of the coin is that it is now becoming increasingly more difficult to navigate the profusion of many disparate clinical terminology standards, which often span millions of concepts.

Objective: The objective of our study was to develop a methodology for integrating large amounts of structured clinical information that is both terminology agnostic and able to capture heterogeneous clinical phenotypes including problems, procedures, medications, and clinical results (such as laboratory tests and clinical observations). In this context, we define phenotyping as the extraction of all clinically relevant features contained in the EHR.

Methods: The scope of the project was framed by the Common Meaningful Use (MU) Dataset terminology standards; the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), RxNorm, the Logical Observation Identifiers Names and Codes (LOINC), the Current Procedural Terminology (CPT), the Health care Common Procedure Coding System (HCPCS), the International Classification of Diseases Ninth Revision Clinical Modification (ICD-9-CM), and the International Classification of Diseases Tenth Revision Clinical Modification (ICD-10-CM). The Unified Medical Language System (UMLS) was used as a mapping layer among the MU ontologies. An extract, load, and transform approach separated original annotations in the EHR from the mapping process and allowed for continuous updates as the terminologies were updated. Additionally, we integrated all terminologies into a single UMLS derived ontology and further optimized it to make the relatively large concept graph manageable.

Results: The initial evaluation was performed with simulated data from the Clinical Avatars project using 100,000 virtual patients undergoing a 90 day, genotype guided, warfarin dosing protocol. This dataset was annotated with standard MU terminologies, loaded, and transformed using the UMLS. We have deployed this methodology to scale in our in-house analytics platform using structured EHR data for 7931 patients (12 million clinical observations) treated at the Froedtert Hospital. A demonstration limited to Clinical Avatars data is available on the Internet using the credentials user “jmirdemo” and password “jmirdemo”.

Conclusions: Despite its inherent complexity, the UMLS can serve as an effective interface terminology for many of the clinical data standards currently used in the health care domain.

(JMIR Med Inform 2014;2(1):e5) doi:10.2196/medinform.3172

KEYWORDS

meaningful use; semantic interoperability; UMLS; SNOMED CT; LOINC; RxNorm; CPT; HCPCS; ICD-9; ICD-10

Introduction

The Definition of Meaningful Use

The Health Information Technology for Economic and Clinical Health Act, enacted as part of the American Recovery and Reinvestment Act of 2009, introduced the concept of Meaningful Use (MU) of information technology in health care. The definition of MU in this context is complex and consists of several objectives and measures that health care providers have to demonstrate in three stages and within strict timelines in order to be eligible for early adopter incentives, and later on to avoid penalties for noncompliance starting in 2015. The MU legislation was designed to transform US health care through the development of processes and standards to capitalize on information in individual medical records and to create data resources that would result in better health care for the greater population.

As part of this process, the legislation mandated the use of standard terminologies for the electronic exchange of health information. In particular, the Office of the National Coordinator for Health Information Technology defined a common set of MU data elements for which certification would be required across a number of electronic health records (EHR) interoperability certification criteria. The EHR interoperability can be further categorized into: (1) foundational, the ability to send information from one system to another, but without the need for interpretation on the receiving end; (2) structural, the syntax, format of simply messaging standards to provide transport of the information; and finally, the most challenging, (3) semantic interoperability, which allows the receiving system to interpret and integrate the received information [1]. The Common MU Dataset has profound consequences for semantic interoperability, as it defines a set of strict terminology standards to be used within a certified EHR. A summary of these is provided in Table 1 and introduced in more detail below.

Biomedical Terminologies

The Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) is one of the most widely used biomedical terminologies in the world. It provides terms, synonyms, and relations covering a number of clinical domains including diseases, findings, and procedures [2]. The Logical Observation Identifiers Names and Codes (LOINC) is a universal standard for identifying laboratory observations. It is considered the *lingua franca* of the clinical observation exchange with its more than 20,000 users in 150 countries [3]. The National Drug Code (NDC) is a well established drug standard that is required in electronic pharmacy claims [4]. The RxNorm is a more recent standardized drug nomenclature designed to facilitate medication reconciliation. It incorporates a number of other drug terminologies, as well as maps to the NDC [5]. The Health care Common Procedure Coding System (HCPCS), maintained by the Centers for Medicare & Medicaid Services (CMS), is a standardized coding system for describing items and services provided in the delivery of health care [6]. It incorporates the Current Procedural Terminology (CPT), a coding system maintained by the American Medical Association, to identify medical services and procedures used by physicians and other health care professionals [7]. The American Dental Association, for accurate reporting of dental treatment [8], developed the Code on Dental Procedures and Nomenclature (CDT). 3M Health Information Systems have developed the International Classification of Diseases Tenth Revision Procedure Coding System (ICD-10-PCS) for the CMS as a replacement for the International Classification of Diseases Ninth Revision Clinical Modification (ICD-9-CM) [9]. The International Classification of Diseases Tenth Revision Clinical Modification (ICD-10-CM) does not contain a procedure classification in contrast to its predecessor ICD-9-CM, and this is where ICD 10 PCS complements ICD-10-CM. The HCPCS, CDT, and ICD-9-CM are used in US electronic transaction claims with planned replacement of the ICD-9-CM by the ICD-10 in October 2014.

US health care relies on a number of different clinical terminology standards with varying levels of overlap and maturity. This already intricate landscape is further complicated by the disparity between billing and MU reporting. For example, SNOMED CT is not allowed in claims reporting and RxNorm combines multiple NDCs under one substance code, rendering detailed package and labeler based billing difficult. The clinical informatics community is now recognizing the need for new tools capable of consuming these heterogeneous resources, hence the term “next generation phenotyping” [10]. In this context, phenotyping is defined as extracting all clinically relevant information from raw EHR data. These clinically relevant features include problems, procedures, medications, and clinical results (such as laboratory tests and clinical observations) annotated with standard clinical terminologies.

Table 1. Common MU Dataset defined in Stage 2 MU Final Rule (Federal Register Vol. 77, No. 171, September 4, 2012) and corresponding vocabulary standards.

Common MU Dataset	Vocabulary standard
1. Patient name	N/A
2. Sex	N/A
3. Date of birth	N/A
4. Race	The OMB[a] Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity, Statistical Policy Directive No. 15, as revised, October 30, 1997
5. Ethnicity	OMB
6. Preferred language	As specified by the Library of Congress, ISO[b]639-2 alpha-3 codes limited to those that also have a corresponding alpha-2 code in ISO 639-1
7. Smoking status	Any of the following SNOMED CT[c] codes- (1) Current every day smoker, 449868002 (2) Current some day smoker, 428041000124106 (3) Former smoker, 8517006 (4) Never smoker, 266919005 (5) Smoker, current status unknown, 77176002 (6) Unknown if ever smoked, 266927001 (7) Heavy tobacco smoker, 428071000124103 (8) Light tobacco smoker, 428061000124105
8. Problems	At a minimum, SNOMED CT International Release July 2012 and US Extension to SNOMED CT March 2012 Release
9. Medications	RxNorm, August 6, 2012 Release
10. Medication allergies	RxNorm, August 6, 2012 Release
11. Laboratory tests	LOINC[d] version 2.40
12. Laboratory values/results	N/A
13. Vital signs (height, weight, BP[e], BMI[f])	N/A
14. Care plan fields including goals and instructions	N/A
15. Procedures	At a minimum, SNOMED CT International Release, July 2012 with US Extension to SNOMED CT March 2012 or the combination of HCPCS[g] and CPT[h] 4 Optional, CDT[i], ICD-10-PCS[j]
16. Care team members	N/A

[a]OMB=Office of Management and Budget

[b]ISO=International Organization for Standardization

[c]SNOMED CT=Systematized Nomenclature of Medicine, Clinical Terms

[d]LOINC=Logical Observation Identifiers Names and Codes

[e]BP=blood pressure

[f]BMI=body mass index

[g]HCPCS=Health care Common Procedure Coding System

[h]CPT=Current Procedural Terminology

[i]CDT=Code on Dental Procedures and Nomenclature

[j]ICD-10-PCS=International Classification of Diseases, Tenth Revision, Procedure Coding System

Local Coding Systems

Many organizations develop their own local coding systems to address these challenges. In fact, to meet the 2014 Edition EHR Certification Criteria, providers are not required to use terminology standards internally as long as they are able to consume them for data portability and clinical quality measures reporting. Convergent Medical Terminology (CMT) is an example of such a solution developed by Kaiser Permanente (KP). CMT serves as the common terminology across all of the KP enterprise, and, at its core is comprised of SNOMED CT, LOINC, and First DataBank drug terminology [11]. However, local coding systems require considerable resources to develop and maintain, and, *ipso facto*, add another layer of complexity to an already convoluted process.

We therefore propose a different solution that relies on the Unified Medical Language System (UMLS) developed and maintained by the National Library of Medicine (NLM) [12]. All of the aforementioned terminology standards are already integrated within the UMLS, which incorporates more than a hundred vocabularies in the biomedical domain. Additionally, the UMLS provides a consistent categorization of all concepts represented in the UMLS Metathesaurus within the UMLS Semantic Network. This makes it an ideal candidate for clinical data integration. While the UMLS has not been designed with a specific intent for bioinformatics, it also incorporates many of the bioinformatics resources, such as the Gene Ontology, the Medical Subject Headings, and the Online Mendelian Inheritance in Man (OMIM), which can further facilitate translational research by bridging clinical informatics and bioinformatics [13].

Significance of This Study

There is now a significant need for integrating patient data from multiple sources, as well as supporting ontology driven querying and reporting on a large scale basis as the transformation of health care from paper to electronic progresses. The UMLS has been widely used as a terminology repository [13,14], in ontology related research [15,16], text mining (via MetaMap) [17], and text processing applications [18]. To our knowledge, with the exception of one proof-of-concept study [19], it has never been actually integrated directly into a clinical workflow as a terminology standard itself. The reasons for this are twofold: (1) the UMLS is technically challenging to work with due to its sheer size and complexity. It encompasses almost three million clinical concepts and eight million synonyms connected by almost 35 million relations (2013AB version). The hardware capabilities to work with such massive terminologies have only recently become available. And (2) before MU, there has been little terminology standardization in the EHR that would warrant an effort to integrate multiple vocabularies. To this day, with the exception of the rather limited coding of insurance claims, many hospital systems still use local coding schemes, which require cumbersome manual translation.

Methods

Data Model

Lightweight object models can rely on ontologies instead of modeling semantics explicitly. We have previously demonstrated this approach in the Observ-OM and VarioML models that were specifically validated for phenotype and genotype information by the GEN2PHEN [20] Consortium [21,22]. At its core, Observ-OM uses only four basic concepts to represent any kind of observation: (1) target, (2) feature, (3) protocol, and (4) value. To this effect, patients become simply collections of observations annotated with clinical terminologies. Each observation has at least one ontology term attached. Overcoding, for example, attaching multiple semantically similar concepts from different vocabularies to a single clinical observation, facilitates information retrieval when code similarity or equivalence have not yet been established in the UMLS. “Hemoglobin; glycosylated (A1C)” (CPT:83036) and “Glucohemoglobin measurement” (SNOMEDCT:40402000) are examples of two such concepts. Multiple terms can also provide additional context, for example, the method used to observe a phenotype (typically with LOINC), while keeping the data model flexible.

Additional semantic information can be derived from the semantic type of the UMLS concept used in the annotation. For example, the concept of Warfarin is typed in the UMLS Semantic Network as a Pharmacologic Substance. Thus, any observation about Warfarin can be inferred to be a medication for the purpose of querying or reporting. Where this is insufficient, we used explicit value sets. For example, the Common MU Dataset defines a set of SNOMED CT terms that together comprise smoking status (see Table 1). In this respect, we also created a custom value set based on the Office of Management and Budget (OMB) standard to represent ethnicity (see Table 2).

Table 2. Clinical Avatars data mapped to the UMLS via MU ontologies.

Clinical Avatars		MU source mapping (direct)	UMLS mapping (automatic)	Term label
Gender		None		
	F		C0015780	Female
	M		C0024554	Male gender
Race		OMB standard		
	African American		C0085756	African American
	Native American		C1515945	American Indian or Alaska Native
	Asian		C0078988	Asians
	White		C0043157	Caucasians
	(no data)		C0086409	Hispanic or Latino
	Pacific Islander		C1513907	Native Hawaiian or other Pacific Islander
	Other/unknown		C1532697	Unknown racial group
Height		LOINC:3137-7	C0365282	Body height measured
Weight		LOINC:3141-9	C0365286	Body weight measured
BSA[a]		LOINC:3139-3	C0365285	Body surface area measured
INR[b]		LOINC:34714-6	C1369580	INR in blood by coagulation assay value
Smoker				
	Y	SNOMED CT:77176002	C0337664	Smoker
	N	SNOMED CT:8392000	C0337672	Nonsmoker
DVT [c]				
	Y	SNOMED CT:128053003	C0149871	Deep venous thrombosis
	N	SNOMED CT:413076004	C1446197	No past history of venous thrombosis
AMI [d]				
	Y	SNOMED CT:57054005	C0155626	Acute myocardial infarction
	N	SNOMED CT:301121007	C0577811	Myocardial perfusion normal
CYP2C9		LNC:46724-1	C1830800	cyp2c9 gene mutations found [identifier] in blood or tissue by molecular genetics method nominal
CYP2C92		LNC:56164-7	C2734139	cyp2c9 gene allele 2 [identifier] in blood by molecular genetics method nominal
CYP2C93		LNC:56165-4	C2734141	cyp2c9 gene allele 3 [identifier] in blood by molecular genetics method nominal
VKORC1		LNC:50722-8	C1978717	vkorc1 gene mutations found [identifier] in blood or tissue by molecular genetics method nominal
VKORC1A		LNC:50722-8	C1978717	vkorc1 gene mutations found [identifier] in blood or tissue by molecular genetics method nominal
VKORC1G		LNC:50722-8	C1978717	vkorc1 gene mutations found [identifier] in blood or tissue by molecular genetics method nominal
Warfarin		RxNorm:11289	C0043031	Warfarin

[a]BSA=body surface area

[b]INR=international normalized ratio

[c]DVT=deep vein thrombosis

[d]AMI=acute myocardial infarction

Terminology Server

The terminology service is built on top of a local database, which is populated with a standard set of vocabularies in the UMLS Active Release (subset of the full release, which includes only the actively updated terminologies). The UMLS is loaded into an Oracle database 11g using the Structured Query Language (SQL) scripts provided with the UMLS distribution and updated in sync with its semiannual release cycle. The currently loaded version is displayed dynamically in the scorecard section of the project home page. As a reference, a 2013AB version set incorporating 89 UMLS terminologies included 2,805,252 unique concepts and 8,622,812 synonyms. RxNorm information is loaded as part of the UMLS distribution, rather than through its own separate release.

For the sake of usability, only a preconfigured subset is displayed as navigable tabs at the top of the browser window (Figure 1 shows this at the top right of the figure). However, all the sources included in the UMLS are potentially browsable and are used in synonym expansion.

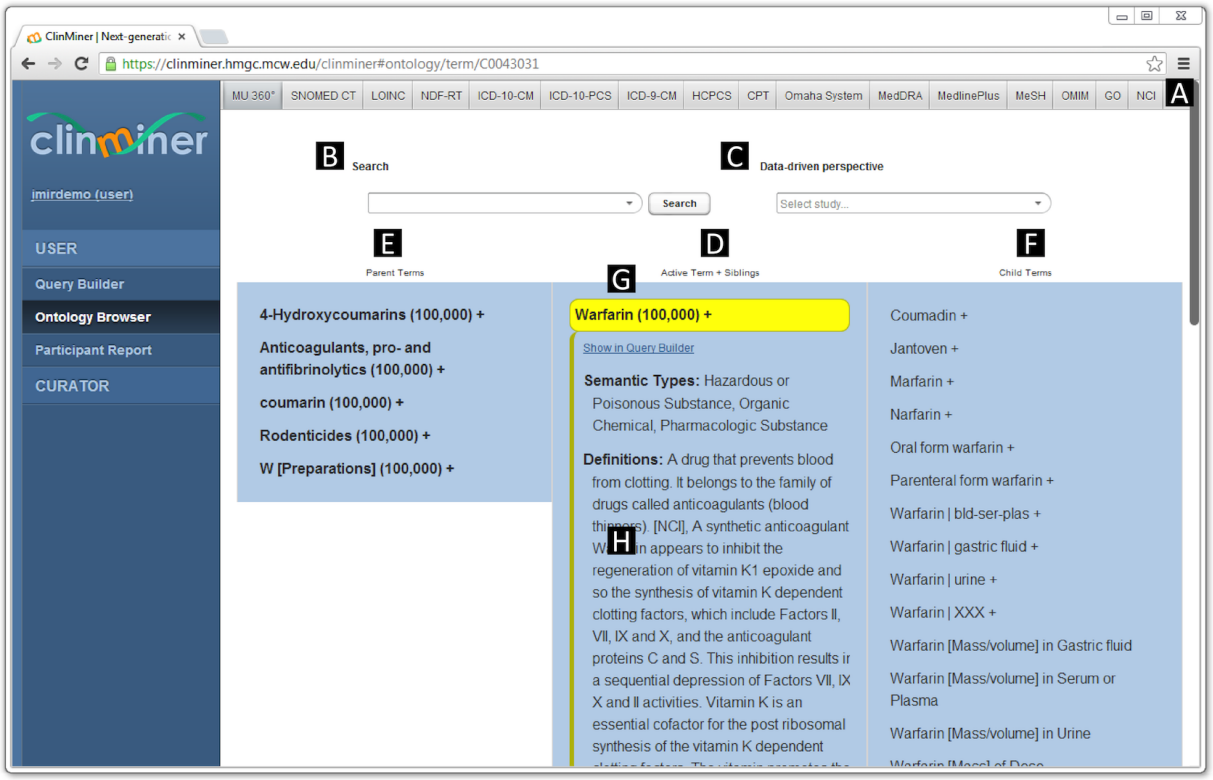
The UMLS comes preconfigured with a broad set of indexes that optimize querying. We use one additional index on top of the attribute value column in the attribute table (MRSAT.ATV) to optimize a dedicated NDC search, which does not exist otherwise as a term code in the concept table (MRCONSO). A label and synonym search was implemented using Oracle Text, a set of Oracle based tools for building text query and document classification applications that provides indexing and text classification capabilities. Individual text tokens are indexed using the term frequency, inverse of the document frequency algorithm, reflecting how often a particular string occurs in the UMLS [23].

Rather than using a complex advanced search interface, we have a single search box that relies on the query relaxation approach (Figure 1). Depending on the context (eg, NDC search requires a different algorithm), the original user query is expanded on the database side into progressively more relaxed versions of the original query. Every search sequence starts with an exact phrase match; then progresses into matching all the tokens in a close proximity (NEAR Procedural Language/Structured Query Language operator); then all words matched (AND) in a phrase; then most words matched (ACCUMulate); and finally falls back to stemming, fuzzy matching and wildcard expansion.

Interpreting a query string using different operator combinations simultaneously allows for a more concise query design. For example, if a user enters a query “rash on examination”, the application can interpret the query in parallel as a single phrase “rash on examination” and “rash” *OR* “on” *OR* “examination” to increase recall. Fuzzy and wildcard matching typically provide the most hits at the expense of precision (a fraction of retrieved instances that are relevant). However, as they are later in the query progression sequence, they are also ranked lower than exact matches, if such exist. For instance, two examples of such fuzzy queries are: (1) “cron disease” (typo in Crohn), which returns the following top three results- “Crohn Disease”, “Crohn's disease”, and “Crohn's disease of large bowel”; and (2) search for “myleoid leukemia” (typo in myeloid), which returns “Myeloid Leukemia”, “Primary Myelofibrosis”, and “Leukemia Myelocytic Acute”.

An example of one of the more powerful features of the Oracle Text search is the ACCUMulate operator that allows parts of the query that did not match to be ignored. That means that it is not necessary to artificially restrain the number of keywords in a query. For example, searching for “cystic fibrosis gene carrier” returns “Cystic fibrosis gene carrier” (all tokens matched), “Carrier of cystic fibrosis gene mutation” (all tokens matched, “of” and “mutation” were ignored), “Encounter due to being a cystic fibrosis carrier” (only “cystic”, “fibrosis”, and “carrier” tokens matched, all others were ignored). In this case, only the first result matched the exact phrase, while the second result had all the keywords, but in a different order, and finally, the last result did not include the keyword “gene”.

Figure 1. Screenshot of ClinMiner's integrated terminology browser. The tabs allow switching between different terminologies and the integrated MU 360 view, default choice (A). Searching. Typing a query into the input field (B) brings up autosuggestions. Selecting a particular string populates the middle panel (D) with search results. Selecting a search result brings back the hierarchical view with the selected term (Warfarin) highlighted in yellow (G). Browsing. Parents of the active term are displayed in the left pane and child terms are displayed in the right pane (F). Meta data for the active term including semantic types, definitions, and non-isa relations to other concepts are displayed in a vignette directly below (H). A plus sign (+) after the term label denotes concepts with children, and the number in brackets reflects the number of participants annotated to a particular term (or its children) in the database. Selecting a study from the drop-down list (C) enables the data driven perspective that displays a compact terminology tree limited to only relevant concepts.



Custom Terminology Browser

The exploration of the UMLS is challenging because of its complexity and lack of obvious starting points, typical of more formal classifications. The UMLS is often displayed as a tree of high level root concepts for the underlying terminologies (cf. the NLM browser provided by the UMLS Terminology Services), but it is in fact more of a tightly interwoven mesh, as it integrates multiple ontology sources with often overlapping coverage and different layers of granularity. Previously, it has been demonstrated that the UMLS is a scale free network that contains both noisy concept hubs (that do not generate meaningful transitive connections, eg, Sudden onset, attribute) and informational concept hubs (that are indispensable for generating useful cross terminology connections, eg, Fever) [24].

Additionally, a graph of UMLS size cannot be effectively analyzed using available state of the art network analysis software, for example, Cytoscape [25]. For this reason, we hypothesized that the more often a particular concept occurs in different sources, the more relevant it is in UMLS navigation. We ranked all the UMLS concepts according to their branching factor (number of children) and number of unique source mappings. The one hundred top ranked concepts were then selected from SNOMED CT, LOINC, and RxNorm separately to achieve equal representation in the final result set of 167 nodes and 230 edges (some concepts overlapped). LOINC codes and parts were considered independently due to their different nature [15]. This smaller network was then plotted in Cytoscape using its hierarchical layout, and 19 identified root concepts formed the entry points for the default MU 360 tab in the ClinMiner terminology browser (Figure 1).

This MU 360 view is a custom UMLS perspective integrating all its sources with a specific focus on MU terminologies. For the purpose of hierarchical display and browsing, we adopted a conservative approach and limited the UMLS traversal to either the UMLS itself, or any of the following terminologies specific to MU- RxNorm, NDF-RT, LOINC, SNOMED CT, HCPCS, and ICD-9-CM. We explicitly ignore hierarchical relations from other terminologies, as in our experience they may add nonsensical paths to query expansion, for example, between "myocardial infarct" and "dermatologic disorders" via "disorder of soft tissue". Additionally, to augment relatively flat LOINC and RxNorm hierarchies, some other relations are treated here as hierarchical, for example "class_of" and "measured_by" in the case of LOINC. This can be seen in Figure 1, where LOINC tests measuring warfarin concentration appear as children of the Warfarin concept.

An example of one of the unique features of mature terminologies such as SNOMED CT in contrast to more simple classification systems such as ICD-9, is that a single concept can exist in multiple places of the hierarchy, for example,

"bronchitis" has two parent terms, "infection" and "bronchial disease". This is difficult to display using a tree like hierarchy, as it requires multiple tree fragments. Instead, the ClinMiner terminology browser displays the active term, all of its parents, siblings, and children terms in three horizontally aligned panes at the same time. The Rat Genome Database originally introduced this approach [26]. When a term is clicked in any of the columns, it becomes the active term and moves to the center column together with its siblings, while adjacent columns update to show parent terms to the left and children terms to the right (Figure 1). This allows for easy exploration of the ontology in both directions, with three levels of terms being visible at all times, and supports multiple inheritance (multiple parents) in a single view.

Extract, Load, and Transform

An automatic process translates original codes in patients' EHR data to their corresponding UMLS concept unique identifiers (CUIs). Figure 2 illustrates an overview of this, and Figure 3 shows the details of the transformation. This is a bidirectional process, as the UMLS codes are also projected back into source terminologies, which can reactivate concepts (when the UMLS CUI is mapped to both active and retired versions of the same concept in the source vocabulary), as well as provide views based on terminologies that were not originally used to annotate the data. For example, in the demonstration it is possible to explore the Clinical Avatars data in the NCI Thesaurus and OMIM tabs (Figure 1), although no direct mappings to the NCI Thesaurus or OMIM were made initially. This is also illustrated in Figure 3, where the original SNOMED CT concept "Deep venous thrombosis" (SNOMED CT:128053003) is translated via the mapped UMLS concept "Deep Vein Thrombosis" (C0149871) into the ICD-10-CM concept "Acute embolism and thrombosis of unspecified deep veins of lower extremity" (ICD10:I82.40). To see the contents of each individual node in this transformation, please see Multimedia Appendix 1.

To facilitate queries across thousands of patients, the transformation process also includes query expansion and complement creation. For example, a patient with "deep vein thrombosis" and "acute myocardial infarction" would, at this step, also be automatically annotated with "cardiovascular diseases", the parent term for these two concepts, as well as negated "No past history of venous thrombosis" and "Myocardial perfusion normal", when no annotations were made to these concepts for this patient. In addition to the intentional restrains on the UMLS traversal described earlier, query expansion is limited to concepts that are within the same UMLS Semantic Network (ie, sharing the same semantic type), as shown in Figure 3.

In terminologies that use multiple inheritance as a design pattern (eg, SNOMED CT vs ICD-9), a single term can exist in multiple paths. Additionally, different granularities and overlap across source terminologies lead to hierarchical cycles (loops). Patient level query expansion adds to this complexity as patients can have multiple annotations of the same type or varying levels of overlapping granularity (see the earlier example of "bronchitis" and "infection"). The simple addition of branch counts would in this case lead to inflated numbers. For this reason, sets of unique patient identifiers have to be propagated across the ontology graph to precalculate accurate patient level counts at every level of the ontological hierarchy, which would eliminate the aforementioned issues and produces a directed acyclic graph. From this, it is straightforward to calculate the propagated negated information as a relative complement of a set of propagated patient terms with respect to all propagated terms across all patients.

In order to minimize the user effort involved in browsing large hierarchies, the ontology graph is additionally approximated as a minimum Steiner tree problem [27]. This produces a more compact reconnected terminology tree, which includes only the concepts that appear in the selected dataset and their best connected parent concepts, rather than all of the potentially available concepts within the UMLS graph. Selecting from the "Data-driven perspective" drop-down in Figure 1 enables this view. This process also identifies orphaned nodes that were otherwise disconnected from hierarchy, placing them at the root of the tree.

Figure 2. An overview of the extract, load, and transform (ELT) process. Data is extracted from multiple sources including disease registries, hospital's EHR system, and clinical notes.

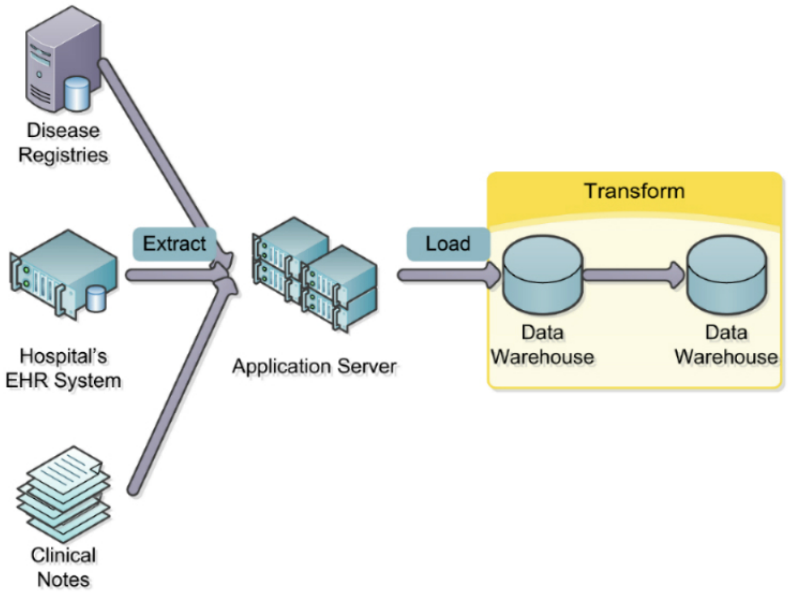
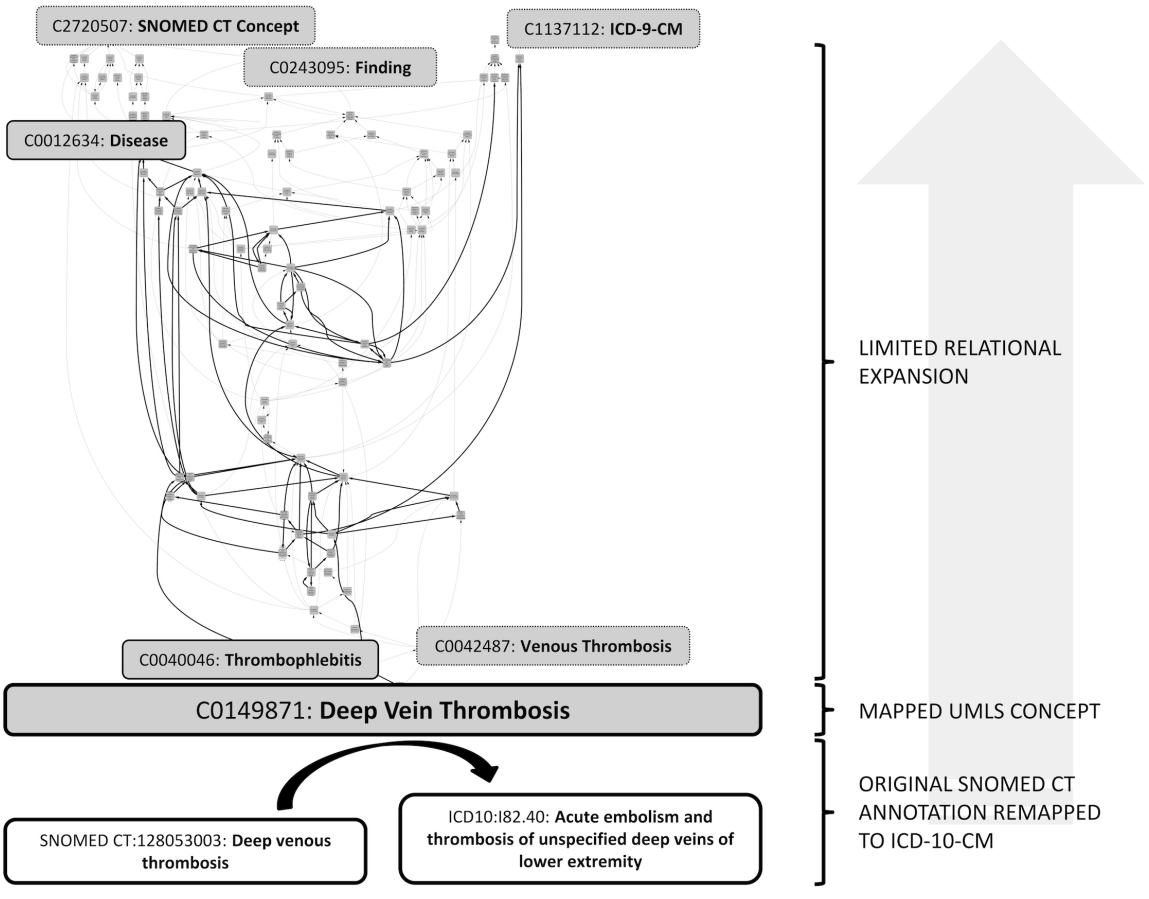


Figure 3. An example of the transformation stage in the extract, load, and transform (ELT) process (for a higher resolution image, see Multimedia Appendix 1). The SNOMED CT annotation "Deep venous thrombosis" made originally in the EHR, is mapped in the UMLS to the "Deep Vein Thrombosis" concept, and can be further remapped into the UMLS source concepts such as the ICD-10-CM "Acute embolism and thrombosis of unspecified deep veins of lower extremity" concept shown in the lower right portion of the figure. The UMLS concept "Deep Vein Thrombosis" is then expanded across a set of parent concepts that are within the same UMLS Semantic Network (solid lines). The concepts characterized by a different semantic type are not included in the expansion (dotted lines). In this example, two parent concepts of "Deep Vein Thrombosis", "Thrombophlebitis and Venous Thrombosis" have semantic types "Disease" or "Syndrome and Pathologic Function" respectively. Thus, the expansion does not include the term "Venous Thrombosis", as the semantic type is different from the originating concept's semantic type ("Disease or Syndrome"), but does include "Thrombophlebitis", which share the same semantic type. There were four high level concepts that were additionally highlighted at the top of the figure, out of which "Disease" is the only one included in the expansion.



Web Front-End

The application was developed in Java using enterprise Java technologies- Spring Framework, Spring Roo, Java Persistence Application Programming Interface (Java Persistence Application Programming Interface, EclipseLink provider), Apache Maven, and AspectJ Vaadin, a Java Web application framework that extends Google Web Toolkit, was used to provide the rich Internet application experience. Apache Tomcat provided the Web container. The Apache Hypertext Transfer Protocol Server isolates the Web container and forces encryption on all browser connections with 256-bit Transport Layer Security.

Data Sharing

Requests for a virtual machine image containing a preconfigured version of ClinMiner can be made to the corresponding author. We also welcome data submissions to our local instance, which can then be securely accessed over the Internet, so there is no need for additional deployment.

Results

Simulated Data

To drive the initial implementation, we used simulated patient data kindly provided by the Clinical Avatars project [28]. The Laboratory for Personalized Medicine created the Clinical Avatars and developed a methodology for creating virtual representations of people for the purpose of conducting personalized medicine simulations. This simulation uses a realistic statistical distribution of patient characteristics such as age, gender, ethnicity, and genotype based warfarin response, and represents a typical set of elements that a researcher would expect in a clinical trial. All avatars data included genotype information on two genes important in warfarin pharmacogenetics: (1) CYP2C9, warfarin metabolizing enzyme; and (2) VKORC1, Vitamin K epOxide Reductase Complex 1. The polymorphisms in these genes are clinically important, as they affect therapeutic warfarin ranges across different racial groups [29]. In this particular case, the dataset used represented a simulation of 100,000 patients (10,836,196 observations) undergoing genotype guided warfarin dosing in the process of initiating oral anticoagulation over 90 days using the Couma Gen protocol [30].

The Clinical Avatars data elements were manually mapped using MU ontologies. The mappings were than validated, and the final set is shown in Table 2. A similar approach is used when annotating real clinical notes, and for this purpose we developed and maintain an internal standard operating procedure. The EHR data has an additional extraction step, where a custom parser strips irrelevant information and encounter based data is transformed into time stamped observations. All preexisting codes in the EHR are loaded *as is*.

Electronic Health Records Data

A “Limited Dataset”, as defined under the Health Insurance Portability and Accountability Act, encompassing 7931 patients was obtained from the Medical College of Wisconsin Clinical Research Data Warehouse for this study. The data extract was in the form of standard Epic Clarity tables for a subset of patients that had an encounter or a problem list in the “Malignant neoplasm of pancreas” (ICD9:157) or “Epilepsy and recurrent seizures” (ICD9:345) code subset. Epic Clarity is an SQL relational database extracted for reporting purposes from Epic Chronicles, the data engine at the heart of Epic’s EHR.

The drug information in the EHR was encoded using Medi-Span terminology, one of the RxNorm sources, which facilitated its automatic translation into the UMLS. The clinical results were encoded as orders using CPT-4 codes or using a fixed category from the “CLARITY_COMPONENT” lookup table. We have manually mapped the top 130 most frequently performed laboratory tests (out of a total of 7766 records in the EPIC “CLARITY_COMPONENT” table) to LOINC, which provided coverage for 94.07% (4,765,012/5,065,315) of all the laboratory tests. The remaining 5.93% (300,303/5,065,315) laboratory tests were left unmapped.

A practical difference between simulated and real EHR data is the much larger concept space, which in this case covered 13,614 unique ICD-9, CPT-4, LOINC, and RxNorm codes. This code set was remapped into the UMLS, which resulted in 13,383 distinct UMLS CUIs, and then expanded as described previously across a limited set of “is_a” and selected other relationships (eg, “has_ingredient”) to facilitate querying, which produced the final set of 30,153 concepts. We have successfully applied this approach in a separate study focused on association rule mining in pancreatic cancer [31].

A demonstration limited to Clinical Avatars data is available on the Internet using the credentials user “jmirdemo” and password “jmirdemo” [32].

Discussion

Extract, Load, and Transform

The crosswalk via the UMLS between different terminologies, as demonstrated in Figure 3, is important for several reasons. Where records are coming from legacy sources, they may use an older coding scheme, for example, ICD-9 or NDC, and this process makes the data browsable via a more expressive terminology, such as SNOMED CT. Additionally, the UMLS transformation alleviates the issue of variability in coding across data sources that use different terminologies, for example, drug information annotated with the Veterans Health Administration National Drug File - Reference Terminology and Medi-Span Master Drug Data Base terminologies can both be reconciled using RxNorm.

The extract, load, and transform approach is substantially different from a more common extract, transform, and load approach, when data is transformed before it is loaded into the data warehouse. Conversely, with extract, load, and transform, we essentially maintain two versions of data: (1) the original annotation set made in the EHR, and (2) a dynamically generated set of UMLS mappings. The original data is never lost, and can be retransformed as new knowledge becomes available.

We are now working on expanding the transformed information to include date and values to support more advanced temporal and value restricted queries. This is a critical step that has a significant impact on the time required to query patient information, however, the actual transformation is relatively resource consuming, for example, it creates 238 annotations per avatar using simulated data and several thousand annotations per patient with real EHR data.

Search and Complexity

A relatively large number of concepts remain unused when annotating clinical data to large terminologies. The UMLS, the largest repository of biomedical terminologies, in its current version spans over 10 million unique concept names from over 160 source vocabularies. Only a subset of the UMLS might be suitable for concept matching [33], and SNOMED CT alone may be enough to represent most of the terms commonly used in medical problem lists [34]. In this study, a cohort of eight thousand patients required between ten and thirty thousand (with query expansion) concepts to capture all clinically relevant features.

While physicians rarely have to deal with ontology hierarchies directly, these are indispensable in clinical research to facilitate query expansion, building transitive closures, and data validation and reconciliation. Any sufficiently large terminology is likely to suffer from some inconsistencies and these, however minor, present unique challenges for ontology end users when they have no direct control over the terminologies they are using. With hundreds of thousands of concepts, traditional navigation through terminology hierarchies becomes impractical. This is why we put a special focus on enhancing search capabilities as well as providing data-driven perspectives that dynamically hide some of this complexity. The search becomes even more important when concepts do not appear where expected or are not in hierarchical relations at all. In our experience, this is the case in approximately one third (data not shown) of LOINC and RxNorm concepts.

Beyond Meaningful Use

Current requirements for terminology standards are not necessarily intuitive and are likely to cause confusion among implementers and subsequent interoperability issues. Optionality for some of the vocabulary standards only adds to the confusion. Existing studies suggest that there is a wide variation in accuracy of MU electronic reporting [35]. Even within a single terminology providers can significantly differ in which code they assign to the same observation [36]. While there are numerous challenges to data capture, the community can best address them through standardization and convergence on key data elements [37].

Interestingly, several resources in the genotype to phenotype space now actively map to the UMLS directly: (1) Orphanet, a portal for rare diseases [38]; (2) the Human Phenotype Ontology project, which provides a structured description of human phenotypic abnormalities [39]; and (3) ClinVar, a novel National Center for Biotechnology Information database for clinical genomics [40], are all good examples of resources that rely on the UMLS to integrate clinical features, conditions, genes, and proteins.

The UMLS incorporates decades of experience and consistency represented by the US National Library of Medicine, which in fact already maintains RxNorm, one of the MU terminologies. It is therefore not unfeasible that the UMLS could provide a clearer path to semantic interoperability.

Conclusions

Despite its inherent complexity, the UMLS can serve as an effective interface terminology for many of the clinical data standards currently used in the health care domain.

Acknowledgments

We thank Vincent A Fusaro and Peter J Tonellato from the Laboratory for Personalized Medicine, Harvard Medical School for providing the Clinical Avatars data. We also thank Stacy Zacher, Glenn Bushee, and Bradley Taylor for their help. This project was funded in part by the Advancing a Healthier Wisconsin endowment at the Medical College of Wisconsin, the National Center for Research Resources, the National Center for Advancing Translational Sciences, and the National Institutes of Health, through grant UL1 RR031973. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' Contributions

TA and MS were both responsible for the conception of this work. TA designed and created the prototype. NS extracted and transformed Clinical Avatars data. NS and TA were both responsible for terminology mappings. All of the authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

An example of the transformation stage in the extract, load, and transform (ELT) process (a version of Figure 3 with higher resolution). The contents from each individual node can be viewed here.

[JPG File, 3MB - medinform_v2i1e5_app1.jpg]

References

1. HIMSS definition of interoperability. 2013 URL: http://www.himss.org/library/interoperability-standards/what-is [accessed 2014-02-24] [WebCite Cache ID 6NcOKIyLu]
2. Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: Overview of the development process and project status. Proc AMIA Symp 2001:662-666 [FREE Full text] [Medline: 11825268]
3. Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, et al. Logical observation identifier names and codes (LOINC) database: A public use set of codes and names for electronic reporting of clinical laboratory test results. Clin Chem 1996 Jan;42(1):81-90 [FREE Full text] [Medline: 8565239]
4. Anonymous. National drug code directory. Hospitals 1969 Oct 1;43(19):80-81. [Medline: 4898575]
5. Parrish F, Do N, Bouhaddou O, Warnekar P. Implementation of RxNorm as a terminology mediation standard for exchanging pharmacy medication between federal agencies. AMIA Annu Symp Proc 2006:1057 [FREE Full text] [Medline: 17238676]
6. Finnegan R. HCPCS--outpatient procedures. J Am Med Rec Assoc 1987 Oct;58(10):20-22. [Medline: 10301915]
7. Anonymous. Current procedural terminology (CPT). JAMA 1970 May 4;212(5):873-874. [Medline: 4909256]
8. Anonymous. Uniform code on dental procedures and nomenclature. Council on Dental Care Programs. J Am Dent Assoc 1969 Oct;79(4):814-817. [Medline: 5266069]
9. Averill RF, Mullin RL, Steinbeck BA, Goldfield NI, Grant TM. Development of the ICD-10 Procedure Coding System (ICD-10-PCS). J AHIMA 1998 May;69(5):65-72. [Medline: 10179251]
10. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013 Jan 1;20(1):117-121 [FREE Full text] [doi: 10.1136/amiajnl-2012-001145] [Medline: 22955496]
11. Dolin RH, Mattison JE, Cohn S, Campbell KE, Wiesenthal AM, Hochhalter B, et al. Kaiser Permanente's convergent medical terminology. Stud Health Technol Inform 2004 Jan;107(Pt 1):346-350. [Medline: 15360832]
12. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993 Aug;32(4):281-291. [Medline: 8412823]
13. 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]
14. Plovnick RM, Zeng QT. Reformulation of consumer health queries with professional terminology: A pilot study. J Med Internet Res 2004 Sep 3;6(3):e27 [FREE Full text] [doi: 10.2196/jmir.6.3.e27] [Medline: 15471753]
15. Adamusiak T, Bodenreider O. Quality assurance in LOINC using description logic. AMIA Annu Symp Proc 2012 Jan;2012:1099-1108 [FREE Full text] [Medline: 23304386]
16. Doing-Harris KM, Zeng-Treitler Q. Computer-assisted update of a consumer health vocabulary through mining of social network data. J Med Internet Res 2011;13(2):e37 [FREE Full text] [doi: 10.2196/jmir.1636] [Medline: 21586386]
17. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp 2001:17-21 [FREE Full text] [Medline: 11825149]
18. Leroy G, Endicott JE, Kauchak D, Mouradi O, Just M. User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. J Med Internet Res 2013 Jan;15(7):e144 [FREE Full text] [doi: 10.2196/jmir.2569] [Medline: 23903235]
19. Johnson KB, George EB. The rubber meets the road: Integrating the Unified Medical Language System knowledge source server into the computer-based patient record. Proc AMIA Annu Fall Symp 1997:17-21. [Medline: 9357580]
20. G2P Knowledge Centre. URL: http://gen2phen.org/ [accessed 2014-03-12] [WebCite Cache ID 6O1K9ms4d]
21. Adamusiak T, Parkinson H, Muilu J, Roos E, van der Velde KJ, Thorisson GA, et al. Observ-OM and Observ-TAB: Universal syntax solutions for the integration, search, and exchange of phenotype and genotype information. Hum Mutat 2012 May;33(5):867-873. [doi: 10.1002/humu.22070] [Medline: 22416047]
22. Byrne M, Fokkema IF, Lancaster O, Adamusiak T, Ahonen-Bishopp A, Atlan D, et al. VarioML framework for comprehensive variation data representation and exchange. BMC Bioinformatics 2012;13:254 [FREE Full text] [doi: 10.1186/1471-2105-13-254] [Medline: 23031277]
23. Salton G. Developments in automatic text retrieval. Science 1991 Aug 30;253(5023):974-980. [doi: 10.1126/science.253.5023.974] [Medline: 17775340]
24. Patel CO, Cimino JJ. A scale-free network view of the UMLS to learn terminology translations. Stud Health Technol Inform 2007 Jan;129(Pt 1):689-693. [Medline: 17911805]
25. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003 Nov;13(11):2498-2504 [FREE Full text] [doi: 10.1101/gr.1239303] [Medline: 14597658]
26. Laulederkind SJ, Tutaj M, Shimoyama M, Hayman GT, Lowry TF, Nigam R, et al. Ontology searching and browsing at the Rat Genome Database. Database (Oxford) 2012 Jan;2012:bas016 [FREE Full text] [doi: 10.1093/database/bas016] [Medline: 22434847]
27. Gilbert EN, Pollak HO. Steiner minimal trees. SIAM J Appl Math 1968 Jan 1;16(1):1-29. [doi: 10.1137/0116001]

28. Fusaro VA, Patil P, Chi CL, Contant CF, Tonellato PJ. A systems approach to designing effective clinical trials using simulations. Circulation 2013 Jan 29;127(4):517-526 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.112.123034] [Medline: 23261867]
29. Limdi NA, Beasley TM, Crowley MR, Goldstein JA, Rieder MJ, Flockhart DA, et al. VKORC1 polymorphisms, haplotypes and haplotype groups on warfarin dose among African-Americans and European-Americans. Pharmacogenomics 2008 Oct;9(10):1445-1458 [FREE Full text] [doi: 10.2217/14622416.9.10.1445] [Medline: 18855533]
30. Anderson JL, Horne BD, Stevens SM, Grove AS, Barton S, Nicholas ZP, Couma-Gen Investigators. Randomized trial of genotype-guided versus standard warfarin dosing in patients initiating oral anticoagulation. Circulation 2007 Nov 27;116(22):2563-2570 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.107.737312] [Medline: 17989110]
31. Adamusiak T, Shimoyama M. EHR-based phenome wide association study in pancreatic cancer. AMIA Summits Transl Sci Proc (in press) 2014 (forthcoming).
32. Clinminer Demo. URL: https://clinminer.hmgc.mcw.edu/clinminer [accessed 2014-03-12] [WebCite Cache ID 6O1KUtIxA]
33. Nadkarni P, Chen R, Brandt C. UMLS concept indexing for production databases: A feasibility study. J Am Med Inform Assoc 2001;8(1):80-91 [FREE Full text] [Medline: 11141514]
34. Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, et al. Evaluation of the content coverage of SNOMED CT: Ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc 2006 Jun;81(6):741-748. [doi: 10.4065/81.6.741] [Medline: 16770974]
35. Kern LM, Malhotra S, Barrón Y, Quaresimo J, Dhopeshwarkar R, Pichardo M, et al. Accuracy of electronically reported "meaningful use" clinical quality measures: A cross-sectional study. Ann Intern Med 2013 Jan 15;158(2):77-83. [doi: 10.7326/0003-4819-158-2-201301150-00001] [Medline: 23318309]
36. Lin MC, Vreeman DJ, McDonald CJ, Huff SM. Auditing consistency and usefulness of LOINC use among three large institutions - using version spaces for grouping LOINC codes. J Biomed Inform 2012 Aug;45(4):658-666 [FREE Full text] [doi: 10.1016/j.jbi.2012.01.008] [Medline: 22306382]
37. Dolin RH. Accuracy of electronically reported "meaningful use" clinical quality measures. Ann Intern Med 2013 Jul 2;159(1):73. [doi: 10.7326/0003-4819-159-1-201307020-00017] [Medline: 23817711]
38. Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. Representation of rare diseases in health information systems: The Orphanet approach to serve a wide range of end users. Hum Mutat 2012 May;33(5):803-808. [doi: 10.1002/humu.22078] [Medline: 22422702]
39. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, et al. The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. Nucleic Acids Res 2014 Jan;42(Database issue):D966-D974 [FREE Full text] [doi: 10.1093/nar/gkt1026] [Medline: 24217912]
40. Riggs ER, Wain KE, Riethmaier D, Savage M, Smith-Packard B, Kaminsky EB, et al. Towards a universal clinical genomics database: The 2012 International Standards for Cytogenomic Arrays Consortium Meeting. Hum Mutat 2013 Jun;34(6):915-919. [doi: 10.1002/humu.22306] [Medline: 23463607]

Abbreviations

CDT: Code on Dental Procedures and Nomenclature
CMS: Centers for Medicare & Medicaid Services
CMT: Convergent Medical Terminology
CPT: Current Procedural Terminology
CUI: concept unique identifier
EHR: electronic health record
HCPCS: Health care Common Procedure Coding System
ICD-9-CM: International Classification of Diseases Ninth Revision Clinical Modification
ICD-10-CM: International Classification of Diseases Tenth Revision Clinical Modification
ICD 10 PCS: International Classification of Diseases-10 Procedure Coding System
KP: Kaiser Permanente
LOINC: Logical Observation Identifiers Names and Codes
MU: meaningful use
NDC: National Drug Code
NLM: National Library of Medicine
OMB: Office of Management and Budget
OMIM: Online Mendelian Inheritance in Man
SNOMED CT: Systematized Nomenclature of Medicine, Clinical Terms
SQL: Structured Query Language
UMLS: Unified Medical Language System

Edited by G Eysenbach; submitted 09.12.13; peer-reviewed by C Weng, M Dogan, S Lim Choi Keung; comments to author 11.02.14; revised version received 18.02.14; accepted 23.02.14; published 18.03.14.

Please cite as:
Adamusiak T, Shimoyama N, Shimoyama M
Next Generation Phenotyping Using the Unified Medical Language System
JMIR Med Inform 2014;2(1):e5
URL: http://medinform.jmir.org/2014/1/e5/
doi:10.2196/medinform.3172
PMID:25601137

Original Paper

Assimilation of Web-Based Urgent Stroke Evaluation: A Qualitative Study of Two Networks

Rajendra Singh[1], PhD; Lars Mathiassen[2], PhD; Jeffrey A Switzer[3], DO; Robert J Adams[4], MS, MD

[1]Arnold School of Public Health, Health Services Policy and Management, University of South Carolina, Columbia, SC, United States

[2]Robinson College of Business, Center for Process Innovation, Georgia State University, Atlanta, GA, United States

[3]Medical College of Georgia, Department of Neurology, Georgia Regents University, Augusta, GA, United States

[4]Medical University of South Carolina, Department of Neuroscience, Medical University of South Carolina, Charleston, SC, United States

Corresponding Author:
Rajendra Singh, PhD
Arnold School of Public Health
Health Services Policy and Management
University of South Carolina
915 Greene Street, Discovery 1 - Room 349
Columbia, SC, 29208
United States
Phone: 1 803 777 8133
Fax: 1 803 777 1836
Email: rsingh@mailbox.sc.edu

Abstract

Background: Stroke is a leading cause of death and serious, long-term disability across the world. Urgent stroke care treatment is time-sensitive and requires a stroke-trained neurologist for clinical diagnosis. Rural areas, where neurologists and stroke specialists are lacking, have a high incidence of stroke-related death and disability. By virtually connecting emergency department physicians in rural hospitals to regional medical centers for consultations, specialized Web-based stroke evaluation systems (telestroke) have helped address the challenge of urgent stroke care in underserved communities. However, many rural hospitals that have deployed telestroke have not fully assimilated this technology.

Objective: The objective of this study was to explore potential sources of variations in the utilization of a Web-based telestroke system for urgent stroke evaluation and propose a telestroke assimilation model to improve stroke care performance.

Methods: An exploratory, qualitative case study of two telestroke networks, each comprising an academic stroke center (hub) and connected rural hospitals (spokes), was conducted. Data were collected from 50 semistructured interviews with 40 stakeholders, telestroke usage logs from 32 spokes, site visits, published papers, and reports.

Results: The two networks used identical technology (called Remote Evaluation of Acute isCHemic stroke, REACH) and were of similar size and complexity, but showed large variations in telestroke assimilation across spokes. Several observed hub- and spoke-related characteristics can explain these variations. The hub-related characteristics included telestroke institutionalization into stroke care, resources for the telestroke program, ongoing support for stroke readiness of spokes, telestroke performance monitoring, and continuous telestroke process improvement. The spoke-related characteristics included managerial telestroke championship, stroke center certification, dedicated telestroke coordinator, stroke committee of key stakeholders, local neurological expertise, and continuous telestroke process improvement.

Conclusions: Rural hospitals can improve their stroke readiness with use of telestroke systems. However, they need to integrate the technology into their stroke delivery processes. A telestroke assimilation model may improve stroke care performance.

(JMIR Med Inform 2014;2(1):e6) doi:10.2196/medinform.3028

KEYWORDS

telemedicine; stroke; telestroke; information technology assimilation; case study

Introduction

Stroke is a leading cause of death and serious, long-term disability in the United States. In 2008, nearly 800,000 people suffered a stroke, resulting in the deaths of more than 134,000 people [1]. Stroke-related costs are also very high—in 2007, the estimated mean lifetime costs resulting from stroke in the United States were $140,000 per patient and the estimated total costs were $62.7 billion [2]. Worldwide, 15 million people suffer stroke each year; of these, 5 million die and another 5 million are permanently disabled [3].

For ischemic (ie, nonbleeding) strokes, a blood-clot dissolving drug tissue plasminogen activator (tPA) greatly reduces the risk of severe disabilities if administered within 4 ½ hours from the onset of stroke symptoms [4,5]. However, for nonischemic (ie, hemorrhagic) strokes, the tPA treatment would be fatal to the patient. The clinical diagnosis of stroke is therefore challenging; emergency physicians may have difficulty differentiating an ischemic stroke from conditions with a similar presentation and determining which patients would benefit from tPA. Therefore, urgent stroke diagnosis requires readily available neurological expertise, which puts rural hospitals in the difficult position of either transferring all stroke patients to regional medical centers or acquiring such expertise at the risk of variable demand and negative budget impacts.

Information technology (IT)—in the form of specialized Web-based telemedicine systems that include videoconferencing and supporting applications that enable a remote stroke specialist to view and evaluate a patient—has helped address the challenge of urgent stroke care in underserved communities [6]. Such systems, referred to as telestroke, allow emergency departments (EDs) in hospitals to receive patients with suspected stroke and to quickly determine (after consulting a remote stroke specialist) whether to administer tPA [7,8]. Consequently, rural hospitals can offer patients the same emergency stroke care as larger hospitals, provided they collaborate with the larger hospital through telestroke. Despite these technological advancements, telestroke systems in rural hospitals remain underutilized. This may explain, in part, why systemic treatment of stroke patients with tPA remains very low—reportedly between 3% and 5% nationally [9]. This research examines the postdeployment utilization of telestroke across EDs of participating rural hospitals in 2 telestroke networks. In particular, this research explains variations in utilization of a Web-based telestroke system for urgent stroke evaluation.

IT utilization (or assimilation) can be defined as "the extent to which the use of technology diffuses across the organizational projects or work processes and becomes routinized in the activities of those projects and processes" [10]. Following Cooper and Zmud's [11] six-stage model of IT implementation process, IT assimilation combines routinization (when IT application usage is encouraged as a normal activity) and infusion (when increased organizational effectiveness results from using the IT application to its fullest potential). Before IT assimilation can occur, the organization must already have completed the earlier stages of IT implementation. These stages include initiation (when the organization has scanned its problems, opportunities, and available IT solutions, and found a match between an IT solution and its application), adoption (when the organization has decided to invest resources to implement the IT solution), adaptation (when the IT application has been developed, installed, and made available for use), and acceptance (when organizational members have committed to using the IT application) [11]. Thus, IT assimilation occurs when an organization progresses beyond initial technology deployment and integrates it into day-to-day work processes to enhance business performance [12-14].

Recent studies have explored IT adoption in health care organizations [15-19], but Fichman and Kemerer [20], Zhu et al [13], and others have noted that adoption does not always result in effective assimilation of the technology. Still, relatively few studies have explored IT assimilation in health care organizations. Notable examples include Meyer and Goes' [21] nine-stage model of assimilation of technological innovations in hospitals, Ash's [22] investigation of assimilation ("internal diffusion and infusion") of three technological innovations across 67 academic health science centers, Chau and Hu's [23] study of telemedicine assimilation in hospitals, Leonard and Sittig's [24] IMPROVE-IT model connecting IT utilization to health outcomes, and Davidson and Heslinga's [25] examination of assimilation of electronic health records in physician practices. Despite these and a few other IT assimilation studies in health care organizations, there are no in-depth examinations of variations in assimilation of a particular technology across hospitals.

Recent telestroke literature has focused on the organizational, managerial, financial, technical, and legal issues that influence adoption. The enablers of telestroke adoption include a stroke systems of care model with primary and comprehensive stroke centers of excellence, statewide and local stroke champions, pre-hospital and in-hospital coordination, favorable regulatory and reimbursement policies, stakeholder support and communication, and appropriate IT infrastructure [26-30]. The barriers to telestroke adoption include lack of public awareness of stroke symptoms and the need for timely treatment, logistical and coordinative challenges of providing appropriate and timely treatment, limited availability of local neurologists, physician reluctance to use tPA, regulatory and jurisdictional issues, technical and financial issues, and lack of stakeholder support [29-32]. However, to our knowledge, no studies have explored factors that enable telestroke assimilation (ie, postdeployment utilization) in hospitals. Hence, the aim of this study was to examine potential sources of variations in telestroke assimilation in hospitals that offer urgent stroke evaluation and management in collaboration with a tertiary hospital.

Methods

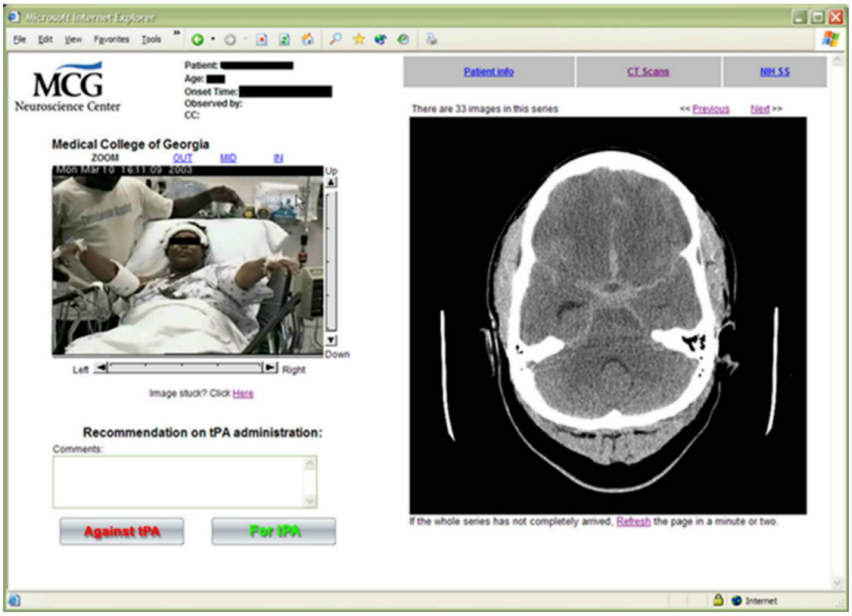
Research Design and Case Context

Based on purposive sampling [33], we organized this research as an exploratory, qualitative case study of 2 stroke networks in Georgia and South Carolina. Each network includes a hub—a comprehensive stroke center at the Georgia Regents University (GRU) and at the Medical University of South Carolina (MUSC)—and connected spokes (ie, rural hospitals supported

by the hub). The two networks use the same technology (Remote Evaluation of Acute isCHemic stroke, REACH), they are of similar size and complexity (17 and 15 spokes, respectively), and they operate in similar contexts (providing services to EDs in rural hospitals in the southeast United States). This design allowed us to conduct cross-case comparisons [33,34] of how hub-related characteristics may influence telestroke assimilation across spokes.

Recognizing the potential of using telestroke to link hub-based specialists to rural hospitals, a team of GRU neurologists developed the REACH system. The system comprised a mobile, Internet-ready REACH cart (with a mounted adjustable camera, a phone, and a high-resolution monitor) that could be wheeled into the ED room where the stroke patient was being examined. As shown in Figure 1, the software embedded within the cart included a Web-based interface to view and share computed tomography (CT) scans and other patient-related information stored within the hospital's electronic medical record system (EMR), picture archiving and communication system, and laboratory information system. In February 2003, GRU signed a contract with the first spoke where it placed a REACH cart. The spoke ED staff activated the REACH system if a patient with suspected stroke arrived within 4 hours of onset of symptoms and then contacted the on-call stroke specialist. The specialist logged onto REACH website via any broadband Internet-connected computer and completed the consultation with a recommendation to administer (or not to administer) tPA to the patient. A for-profit company (REACH Health Inc) provided round-the-clock technology support. By August 2012, 17 hospitals had joined the GRU-REACH network. The MUSC-REACH network was established when one of the founders of REACH joined MUSC and set up a telestroke program in South Carolina in May 2008. By August 2012, 15 hospitals had joined MUSC-REACH. The design, technical details, outcomes, and organizational challenges of REACH have been published elsewhere [26,35-47].

Figure 1. REACH Web interface showing a patient's CT scan.



Data Sources

We collected primary data between March and August 2012 by visiting the 2 hubs and 8 selected spokes (Table 1). These spokes—4 in each network—were selected (out of 32) based on REACH utilization; they included spokes with higher than average and lower than average REACH utilization in the network. During our field visits, we interviewed key stakeholders associated with telestroke, such as administrators, managers, ED physicians, nurses, neurologists, and emergency medical service (EMS) representatives. We asked all respondents to share their experiences of using REACH. The semistructured interviews lasted about 1 hour each. Altogether, we conducted 50 in-person and telephone interviews with 40 stakeholders. To enhance data quality, we collected evidence from multiple sources, including published papers related to the REACH network, as well as internal presentations, emails, and reports. This secondary data helped to gain insight into the current and historical context of REACH implementation in the two networks, and to validate the information collected during the interviews.

We also collected archival data from the 2 hubs related to REACH consultations with each spoke since the start of the telestroke program. To account for variations in spoke ED volume across hospitals, we adjusted the annual rate of REACH consultations at each hospital by its reported ED volume. We refer to the average adjusted annual telestroke consultation rate (calculated as number of REACH consultations/year per 10^4 ED volume) as REACH assimilation. Thus, we consider the REACH-enabled consultation rate as a proxy for telestroke assimilation. It must be emphasized that this paper focuses on the decision-making enabled by the telestroke technology; therefore, we have examined REACH consultations rather than the resulting tPA usage.

Table 1. Primary and secondary data sources.

Primary data sources	Secondary data sources
15 semistructured interviews at 2 hubs (with neurologists, stroke coordinators, ED nurse managers, stroke service line manager, and data analyst)	14 published papers [26,35-47] 10 internal documents related to 2 hubs (including internal presentations, emails, reports, and meeting notes) Archival data related to REACH consultations with each spoke
30 semistructured interviews at 8 spokes (with chief executive officers and chief operations officers, stroke coordinators, neurologists, ED directors, ED physicians, ED nurses, quality managers, radiology nurses, and EMS directors) One staff meeting at a spoke	15 internal documents related to 8 spokes (including presentations, stroke protocols, emails, and meeting notes)
5 semistructured interviews at REACH Health Inc (with chief executive officer, chief technology officer, marketing director, business manager, and IT specialist) One REACH system demonstration	5 internal documents (including presentations, technical specifications, and meeting notes)

Results

Network-Level Variation in Telestroke Assimilation

Table 2 shows basic information about the spokes. The 17 spokes in GRU-REACH network have 1831 beds (range 10-236, mean 108, SD 76) and receive more than 300,000 ED patients/year. Between February 2003 (when GRU-1 became a spoke) and August 2012 (when we collected the data), these spokes reported 2179 REACH consultations (range 48-280, mean 128, SD 71). The 15 spokes in MUSC-REACH network have 2482 beds (range 25-453, mean 165, SD 122) and receive more than 450,000 ED patients/year. Between May 2008 and August 2012, these spokes reported 2753 REACH-enabled consultations (range 60-411, mean 183, SD 107).

Figure 2 compares the REACH assimilation across spokes in the two networks. Except for 1 spoke (MUSC-4 in Table 2 rarely used telestroke and left the network in November 2010 after hiring a neurologist), the MUSC-REACH network outperformed GRU-REACH with a 35% higher REACH assimilation (24.32 vs 18.01; *P*=.07). One reason is that when one of REACH's founding neurologists joined MUSC, he leveraged the lessons learned during the development of the GRU network. This neurologist explained: "When I started the MUSC telestroke program, I did not want to make the same mistakes we did when we developed the Georgia REACH program."

Figure 2. Variation in telestroke assimilation across networks.

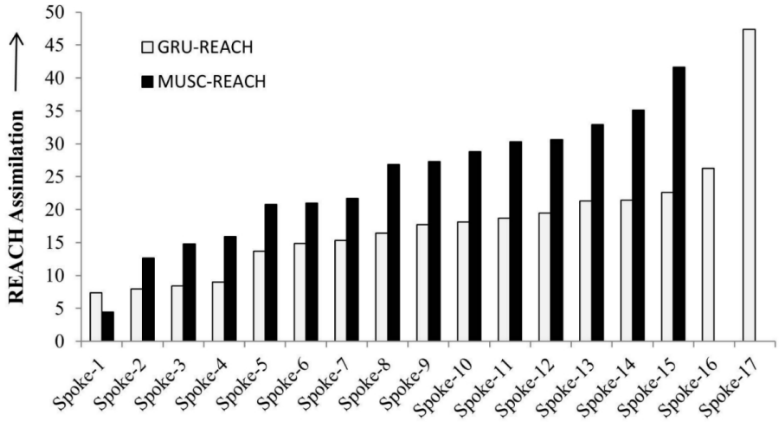


Table 2. Network characteristics and REACH assimilation data.

Telestroke network	Spoke hospital	Joining date	No. of beds	Primary stroke center	Stroke coordinator	Local neurologist	REACH assimilation[a]
GRU	1	2/1/03	72	No	No	No	18.70
GRU	2	3/1/03	47	No	No	No	15.31
GRU	3	7/1/03	50	No	No	No	26.24
GRU	4	8/1/03	10	No	No	No	18.12
GRU[b]	5	9/1/03	56	No	No	No	21.33
GRU	6	3/1/04	65	No	No	No	16.44
GRU	7	4/1/04	20	No	No	No	9.01
GRU[b]	8	2/1/05	52	No	No	No	17.71
GRU	9	3/1/06	71	No	No	No	13.72
GRU	10	1/1/08	191	No	No	Yes	7.93
GRU	11	8/1/08	236	Yes	No	Yes	7.35
GRU	12	6/1/09	40	No	No	No	19.51
GRU	13	10/1/09	190	Yes	Yes	Yes	21.42
GRU[b]	14	10/1/09	196	Yes	Yes	Yes	22.65
GRU	15	1/1/10	180	Yes	Yes	Yes	47.40
GRU	16	3/1/10	163	No	No	Yes	8.39
GRU[b]	17	11/1/10	192	No	Yes	Yes	14.88
MUSC	1	5/1/08	131	No	No	Yes	20.81
MUSC	2	5/6/08	140	No	No	Yes	30.66
MUSC[b]	3	5/7/08	453	No	Yes	Yes	14.77
MUSC	4	9/1/08	220	No	No	No	4.47
MUSC	5	9/18/08	124	No	No	No	15.89
MUSC	6	12/23/08	25	No	Yes	No	35.10
MUSC[b]	7	1/20/10	45	Yes	Yes	Yes	41.62
MUSC[b]	8	3/26/10	288	Yes	Yes	Yes	26.84
MUSC[b]	9	5/19/10	121	No	Yes	Yes	32.93
MUSC	10	7/29/10	79	No	No	No	28.80
MUSC	11	8/26/10	231	No	No	No	21.69
MUSC	12	01/21/11	116	No	No	Yes	12.64
MUSC	13	2/28/11	105	No	No	Yes	30.30
MUSC	14	2/28/11	50	No	No	Yes	21.00
MUSC	15	3/2/11	354	No	Yes	Yes	27.30

[a]REACH assimilation calculated as number of telestroke consultations/year per 10^4 ED volume.

[b]Spokes selected for detailed examination (through field visits).

Hub-Level Variation in Telestroke Assimilation

Based on primary and secondary data analysis, we identified several hub-related practices that can explain the superior telestroke assimilation in the MUSC-REACH network. Table 3 presents these findings. These practices include telestroke institutionalization into stroke care, providing resources for telestroke program, support for stroke readiness of spokes, telestroke performance monitoring, and continuous telestroke process improvement.

Table 3. Comparison of hub-level practices.

GRU-REACH hub	MUSC-REACH hub
GRU-REACH hub invited most of the early spokes to become part of the network and subsidized their participation; most recent spokes sought membership without subsidies.	MUSC-REACH hub invited most of the early spokes to become part of the network, but participation was not subsidized; most recent spokes also sought membership without subsidies.
GRU administration considers telestroke as an ongoing experiment rooted in the vision and goodwill of the stroke specialists who developed REACH. As such, the specialists feel REACH is "taken for granted." GRU administration does not provide support for telestroke operations.	MUSC administration considers telestroke an integral part of their neuroscience service line, and therefore provides ongoing support (including director's pay, advertising budget, and administrative salary support for credentialing, billing, operations, and project management).
There is broad consensus among the hub stroke specialists that network performance would benefit from a full-time telestroke coordinator.	A dedicated telestroke coordinator at the hub has been part of the network from the start. She facilitates coordination and training of the spokes' ED staff.
The hub has no established processes for reinforcing telestroke use and related routines at the spokes. There are no continuous quality improvement processes in place. Any problems related to stroke consultations are reported to REACH Health Inc with variable follow-up.	The hub has established processes for reinforcing telestroke use and related routines at the spokes. It has a formal continuous quality improvement process in place. Any problem during telestroke consultation is reported to REACH Health Inc and its resolution is coordinated by the hub staff.
The hub collects spokes' telestroke use data, but there is no systematic analysis of the data.	The hub telestroke coordinator collects spokes' usage data and conducts systematic analysis.
A hub stroke specialist visits spokes when they go live with REACH and at rare occasions for major upgrades. However, there are no ongoing training and follow-up procedures.	A hub telestroke specialist visits spokes when they go live with REACH and maintains regular communication (with some visits) to spokes to understand concerns and train ED staff.
The hub stroke specialists rarely conduct ongoing training for spokes.	The hub facilitates occasional breakfast meetings, lunch-and-learn, mock-consults, and dinners with spoke ED physicians and nurses to discuss issues.
The hub has no formal system to provide site-specific feedback.	The hub provides site-specific performance data. As an MUSC-REACH stroke specialist told us, "The sites love to receive such feedback."

Spoke-Level Variation in Telestroke Assimilation

Identifying Characteristics that Explain Spoke-Level Variation

Spoke-level REACH assimilation varied from 7.35 in GRU-11 to 47.4 in GRU-15 (average 18.00), and from 4.47 in MUSC-4 to 41.62 in MUSC-7 (average 24.32). We cannot explain these large variations by length of relationship with the hub or size of the spoke. For example, GRU-3 and GRU-4 joined the network within 1 month of each other, but still showed variation in assimilation (26.24 and 18.12). Moreover, GRU-13 joined the network more than 6 years after GRU-5 and both showed similar REACH assimilation (21.42 and 21.33). Furthermore, MUSC-5 and MUSC-9 had similar number of beds (124 and 121), but showed considerable variation in REACH assimilation (15.89 and 32.93).

To explain the observed variations across all spokes, we first considered the availability of local neurological expertise for post-tPA patient supervision. Seven spokes in GRU-REACH and 10 spokes in MUSC-REACH had an on-call local neurologist. As Table 4 shows, when local neurology support was available, GRU-REACH spokes showed similar assimilation (18.57 vs 17.61, P=.87), whereas MUSC-REACH spokes showed relatively higher assimilation (25.89 vs 21.19, P=.46). Overall, availability of local neurological expertise was associated with a 21.70% improvement in assimilation (22.88 vs 18.80, P=.24). Although these variations do not show statistical significance (the very small sample sizes may explain the P values generated), the data suggest that ED staff sought more telestroke consultations when a neurologist was readily available.

Next, we considered whether stroke center certification had an impact on telestroke assimilation. The US Joint Commission certifies acute care hospitals as "Primary Stroke Centers" if they have specialized knowledge and infrastructure to treat stroke patients. The certification signifies that a hospital has necessary stroke-related facilities (such as ED, EMS, and stroke unit), services (such as neurological, neuro-imaging, laboratory, and clinical support), personnel (such as acute stroke teams), practices (such as written care protocols, outcome and quality improvement activities, and continuing medical education), and commitment and support of the medical organization [48]. Overall, 4 spokes in GRU-REACH and 2 spokes in MUSC-REACH had stroke certification. As Table 4 shows, REACH assimilation was higher in these cases (54.86% higher in GRU-REACH, P=.38; 50.13% higher in MUSC-REACH, P=.38; and 43.93% higher overall, P=.22). Thus, the data suggest that stroke care certification resulted in higher assimilation (again, the very small sample sizes may explain the P values generated).

We also considered the impact of a telestroke coordinator. Such a position may help spokes establish standard processes for stroke care; collect, analyze, and use performance data to continually improve care delivery; and, become a stroke champion in the hospital and in the local community. Four spokes in GRU-REACH and 6 spokes in MUSC-REACH had a dedicated telestroke coordinator. As Table 4 shows, REACH assimilation in the spokes with stroke coordinator was significantly higher than without the coordinator (73.00% higher

in GRU-REACH, *P*=.22; 43.84% higher in MUSC-REACH, *P*=.08; and 62.34% higher overall, *P*=.01), suggesting that having a dedicated coordinator resulted in higher assimilation.

To confirm and elaborate these explanations, we conducted an in-depth analysis of telestroke use at 4 selected spokes in each network. Helped by long-standing relationships with the 2 hubs, we visited these spokes and interviewed key stakeholders associated with stroke operations. These interviews provided additional insights into the current and historical context of REACH implementation at these spokes. Accordingly, we identified several notable practices that can further explain variations in telestroke assimilation across spokes.

Table 4. Impact of spoke characteristics on telestroke assimilation.

Spoke characteristic		REACH assimilation[a] in GRU-REACH	REACH assimilation in MUSC-REACH	Overall REACH assimilation
Local neurological expertise				
	No local neurologist	17.61	21.19	18.80
	Local neurologist	18.57	25.89	22.88
	Difference (%)	5.45	22.18	21.70
Stroke center certification				
	No stroke certification	15.95	22.80	19.37
	Stroke certification	24.70	34.23	27.88
	Difference (%)	54.86	50.13	43.93
Dedicated stroke coordinator				
	No stroke coordinator	15.37	20.69	17.55
	Stroke coordinator	26.59	29.76	28.49
	Difference (%)	73.00	43.84	62.34

[a]REACH assimilation calculated as number of telestroke consultations/year per 10^4 ED volume.

Local Neurological Expertise

In 6 of the 8 spokes that we visited, a combination of local neurological expertise and telestroke provided urgent stroke care. The local neurologists would follow up on patients admitted locally, including post-tPA stroke patients, in the intensive care unit (ICU). In some cases (eg, GRU-5 and MUSC-9), all emergency consultations were handled via telestroke. In other cases, local neurologists also provided acute stroke coverage in the ED either during daytime (GRU-14) or 15 days/month (GRU-17). Overall, the combination of local neurology support and REACH coverage afforded spokes expanded stroke care capability.

Stroke Center Certification

Three of the 8 spokes that we visited (GRU-14, MUSC-7, and MUSC-8) had received primary stroke center certification. This suggests that they had established the necessary infrastructure, acquired stroke-related specialized knowledge, and developed standardized protocols and best practices to manage urgent stroke patients. In 2004, GRU-14 became the first spoke in Georgia to receive certification. To achieve that, GRU-14 set up a dedicated stroke unit, hired three neurologists, and developed standardized protocols (such as a written "stroke code"). Prior to the stroke certification, the local EMS "dreaded bringing stroke patients to the hospital because they were not sure that the hospital had capability to deliver urgent stroke care," and instead took the patients directly to the nearest tertiary medical center. However, as GRU-14 advertised its stroke care capabilities, the local EMS started to bring stroke patients to the hospital. Similarly, after MUSC-7 gained certification in 2010, its acceptance as the preferred stroke care center in the region increased, resulting in a growing number of stroke patients admitted at the hospital. When needed, the ED staff at these hospitals connected to GRU-hub via REACH for consultations.

Dedicated Telestroke Coordinator

At 6 of the 8 spokes that we visited (GRU-14, GRU-17, MUSC-3, MUSC-7, MUSC-8, and MUSC-9), a telestroke coordinator set up and developed requisite processes, and facilitated collaboration within the hospital and with the hub. The coordinator provided ongoing feedback and training to ED nurses to reinforce and improve stroke-related processes, and conducted systematic spoke performance analysis. The coordinator helped to develop best practices (such as taking blood samples for laboratory analysis while the patient was in the CT scan room), which helped to reduce delays in stroke treatment. A full-time coordinator at GRU-17 reviewed each stroke case and reported any deficiencies (eg, missed stroke diagnosis, or delays in CT scan). Spokes (eg, GRU-8) that had no dedicated telestroke coordinator used the services of a part-time coordinator. At GRU-5, MUSC-7, MUSC-8, and MUSC-9, the coordinator conducted community awareness initiatives (including health fairs, and advertisements in the local newspapers, radio, and television) to provide information about stroke symptoms and related services available at the hospital.

Managerial Telestroke Championship

Senior leadership support was critical to establishing and fostering telestroke capability at the spokes. In 5 of the 8 spokes

that we visited (GRU-14, GRU-17, MUSC-7, MUSC-8, and MUSC-9), the senior leadership realized the value of telestroke and encouraged the ED and other staff to make it an integral part of urgent stroke care. They also provided requisite IT infrastructure and resources, and facilitated a culture of continuous improvement. In contrast, at GRU-8, several years of managerial neglect had led to a situation where the ED staff routinely referred stroke patients to other hospitals. Over time, they lost their stroke-handling skills. A nurse manager elaborated on the situation:

> *A few years ago, the ED staff knew what to do in case of a stroke patient. Now, I am not sure they do. I guess they don't know when to trigger the REACH system.*

Stroke Committee of Key Stakeholders

A stroke committee—consisting of a telestroke coordinator, ED physicians and nurses, radiology staff, and EMS—proved essential to improving telestroke practices at GRU-14, GRU-17, MUSC-3, MUSC-7, and MUSC-8. Emphasizing the need for coordination, the chief of medical staff at MUSC-7 said, "We consider stroke to be a team event." In some spokes, the committee also facilitated a cultural change. An ED physician at GRU-14 explained:

> *When I arrived here 3 years ago, we did not have a stroke care culture. The stroke committee took ownership of the stroke program and led the change in culture from within. Now, stroke is a source of identity for the hospital.*

The committees met regularly to discuss issues and to find ways to enhance stroke readiness. The role of ED physicians and nurses in stroke committees was critical. In some spokes (eg, GRU-5), the nurses encouraged the ED physicians to initiate the REACH call, while at others (eg, GRU-14), the ED physicians themselves contacted the remote specialist. In all cases, however, the ED physicians made a decision (to treat locally or transfer patients) based on availability of local neurology support and neuro-ICU facilities in their hospital. Deliberate engagement of the local EMS in some spokes (MUSC-7 and GRU-14) improved stroke performance by reducing patient transportation time. Similarly, a pro-active EMS became an integral part of stroke care at MUSC-8. At MUSC-9, the hospital-owned EMS became the "voice of the hospital."

Continuous Telestroke Process Improvement

Spokes with superior stroke performance (eg, GRU-14 and MUSC-7) focused on improving their stroke delivery processes. Their stroke committees had developed protocols and training procedures to sustain and improve urgent stroke care. Stroke care-related staff at GRU-14 and GRU-17 regularly exchanged best practices and updates with colleagues in other hospitals. The chief financial officer at GRU-17, who trained as a Six Sigma Master Black Belt, had initiated several quality improvement initiatives to improve stroke care. Over time, GRU-17 fostered shared responsibility for stroke care and created a systematic basis for continuous improvements. In contrast, GRU-8 did not have established routines or process improvement initiatives to develop their urgent stroke care capability. At GRU-14, MUSC-7, and MUSC-8, the process improvement initiatives helped achieve the coveted primary stroke center accreditation.

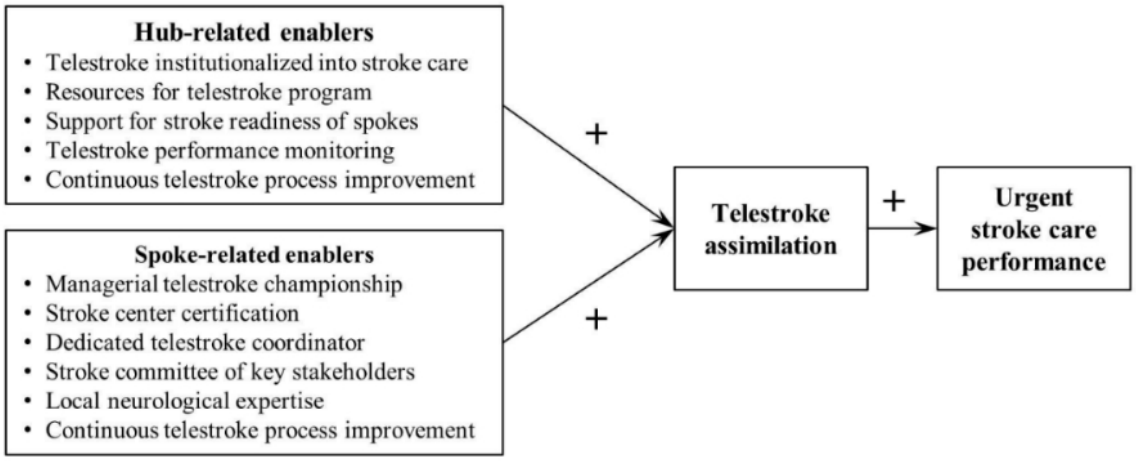
Discussion

Principal Results

The existing IT literature emphasizes how organizational factors enable technology utilization in key processes to enhance business performance [12-14]. Based on this general logic, our study highlights the organizational factors that drive telestroke assimilation at hub and spoke levels. Using data from 2 telestroke networks that operated in similar contexts and relied on the same technology, we investigated the variations in technology assimilation across spokes and zoomed in on organizational factors that could explain this variation.

The identified hub factors included (1) institutionalization of telestroke by making the technology an integral part of stroke delivery, (2) providing required resources for telestroke program, (3) ongoing support for stroke readiness of spokes, (4) telestroke performance monitoring with site-specific feedback, and (5) continuous process improvement to improve telestroke delivery. Similarly, the identified spoke factors included (1) managerial telestroke championship, (2) stroke center certification, (3) dedicated telestroke coordinator, (4) stroke committee consisting of key stakeholders, (5) availability of local neurological expertise, and (6) continuous telestroke process improvement. These empirical findings suggest a telestroke assimilation model (Figure 3) in which specific hub and spoke factors enable increased use of telestroke technology for urgent stroke evaluation. Moreover, as several studies have established, improved urgent stroke evaluation and management—through tPA administration in ischemic strokes or neurosurgical interventions, as appropriate—greatly reduce the chance of severe disabilities [49]. Therefore, the proposed model includes urgent stroke care performance as the overall outcome.

Figure 3. Proposed model of telestroke assimilation.



Comparison With Prior Work

Existing telestroke studies support some of our findings and related elements of the assimilation model. On the hub level, Cho et al [42] found the enabling effect of "institutionalization of telestroke into routine stroke delivery." Similarly, considering "resources for telestroke," Gogan and Garfield [50] found that effective deployment of organizational resources is critical to developing and improving telestroke services. However, few studies have so far examined hub-related organizational factors, such as telestroke process improvement, support for stroke readiness of spokes, and spoke-specific programs for performance monitoring.

On the spoke level, Rogove et al [30] found that lack of leadership support was a major barrier to telestroke, thus emphasizing the enabling role of "managerial telestroke championship." O'Toole Jr. et al [29] identified the lack of "local neurological expertise" in rural areas as a major barrier to telestroke adoption and implementation. Other studies have pointed to the need for "continuous telestroke process improvement." For example, Medeiros de Bustos et al [51] identified the lack of predefined procedures and uneven standards of evaluating stroke care quality as major challenges to telestroke utilization, and Gogan and Garfield [50] identified the need to create appropriate checklists and protocols for stroke care and to engage users in developing repeatable processes. Interestingly, although many studies point to the general need for internal and external coordination for stroke care [29,52], few studies have examined the role of a stroke committee of key stakeholders in directing such efforts or of a dedicated telestroke coordinator in facilitating day-to-day stroke delivery.

Thus, our findings add to the literature in a number of ways. To our knowledge, this is the first study to focus on telestroke assimilation as a key activity in determining how technology contributes to urgent stroke care performance. Second, we have distinguished between hub- and spoke-level factors as the key organizational antecedents to telestroke assimilation. This is particularly important because most studies have focused on spoke-related factors. Finally, we have leveraged our empirical findings to propose a comprehensive model of telestroke assimilation in hospitals that have already deployed the technology.

Limitations

An important limitation relates to the scope of this study. We considered the impact of organizational factors on telestroke assimilation, but did not explore policy-related (eg, reimbursement and incentive structures), technology-related (eg, reliability, ease-of-use, broadband connectivity, and level of integration of telestroke with other IT in the hospital), or behavioral factors (such as physician attitudes toward thrombolysis and technology, and local neurologists' buy-in). Furthermore, we assumed that the patient population characteristics were similar across the spoke hospitals' service areas. It is also important to note that not all hospitals may have the financial resources to hire a neurologist or a dedicated stroke coordinator (which may explain their reluctance or inability to use telestroke). Our findings draw on a comparative case study of two telestroke networks involving a particular technology. Although a case study design has limited generalizability [33,34], it has the advantages of attention to organizational context, dynamics, and multiple stakeholder perspectives [53]. Accordingly, we have provided a rich description of the two networks to help researchers assess and transfer the findings to other settings [54]. We triangulated across data sources, checked against "hard facts" (eg, published documents), used multiple investigators, and iteratively sought feedback on our interpretations from key stakeholders [33,34]. This approach improves the study's confirmability and credibility [54,55]. Finally, the *P* values reported in the results section need to be viewed in light of the low sample size, which affects statistical power and our ability to make meaningful inferences.

Directions for Future Research

Our study suggests some future research directions. First, researchers can validate and improve the proposed telestroke assimilation model by considering additional factors (eg, policy-related, technological, and behavioral) across different networks. Second, researchers can adapt the model to examine postdeployment utilization of telemedicine and other IT (such as EMR and health information exchanges) in health care

organizations. Third, the literature provides several examples of maturity models for IT adoption and assimilation. The term "maturity" relates to the degree of repeatability and optimization of processes, from ad hoc practices, to formally defined steps, to managed result metrics, to active optimization of processes [56]. Accordingly, researchers can leverage our findings to develop a stroke capability maturity model to assess a hospital's current practices and to develop strategies to improve stroke care capability. Finally, researchers can identify and characterize the processes through which health care providers learn to co-create value through collaborative forms of IT.

Conclusions

EDs in rural hospitals with limited neurological expertise face significant challenges in evaluating patients with stroke symptoms. These hospitals need to either transfer stroke patients to larger regional medical centers or hire local neurologists. Recent telemedicine innovations have enabled rural hospitals to connect virtually to regional medical centers for urgent stroke evaluation. However, many hospitals that have deployed telestroke have not assimilated the technology, that is, they have not integrated it into their regular stroke delivery processes. Consequently, neurologic expertise is not used optimally, opportunities for tPA administration may be lost, and patients are transferred out unnecessarily. Based on a detailed examination of variations in telestroke assimilation across two networks, this exploratory research proposes a telestroke assimilation model that includes specific hub- and spoke-related characteristics that can potentially increase IT assimilation by spokes and lead to improved stroke readiness.

Acknowledgments

This research was funded by Genentech.

Conflicts of Interest

JAS has served as a consultant for Genentech and REACH Health Inc and RJA is a cofounder of REACH Health Inc and has cofounder's equity.

References

1. Roger VL, Go AS, Lloyd-Jones DM, Benjamin EJ, Berry JD, Borden WB, et al. Heart Disease and Stroke Statistics--2012 Update: A Report From the American Heart Association. Circulation 2011 Dec 15;125(1):e2-e220. [doi: 10.1161/CIR.0b013e31823ac046]
2. Rosamond W, Flegal K, Furie K, Go A, Greenlund K, Haase N, American Heart Association Statistics CommitteeStroke Statistics Subcommittee. Heart disease and stroke statistics--2008 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Circulation 2008 Jan 29;117(4):e25-146 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.107.187998] [Medline: 18086926]
3. Kuper H, Adami HO, Theorell T, Weiderpass E. The socioeconomic gradient in the incidence of stroke: a prospective study in middle-aged women in Sweden. Stroke 2007 Jan;38(1):27-33 [FREE Full text] [doi: 10.1161/01.STR.0000251805.47370.91] [Medline: 17138948]
4. National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. New Engl J Med 1995;333(24):1581-1588. [doi: 10.1056/NEJM199512143332401] [Medline: 10.1056/NEJM199512143332401]
5. Hacke W, Kaste M, Bluhmki E, Brozman M, Dávalos A, Guidetti D, ECASS Investigators. Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke. N Engl J Med 2008 Sep 25;359(13):1317-1329. [doi: 10.1056/NEJMoa0804656] [Medline: 18815396]
6. Agarwal S, Warburton EA. Teleneurology: is it really at a distance? J Neurol 2011 Jun;258(6):971-981. [doi: 10.1007/s00415-011-5920-5] [Medline: 21311906]
7. Schwamm LH, Holloway RG, Amarenco P, Audebert HJ, Bakas T, Chumbler NR, American Heart Association Stroke Council, Interdisciplinary Council on Peripheral Vascular Disease. A review of the evidence for the use of telemedicine within stroke systems of care: a scientific statement from the American Heart Association/American Stroke Association. Stroke 2009 Jul;40(7):2616-2634 [FREE Full text] [doi: 10.1161/STROKEAHA.109.192360] [Medline: 19423852]
8. Parra C, Jódar-Sánchez F, Jiménez-Hernández MD, Vigil E, Palomino-García A, Moniche-Álvarez F, et al. Development, implementation, and evaluation of a telemedicine service for the treatment of acute stroke patients: teleStroke. Interact J Med Res 2012;1(2):e15 [FREE Full text] [doi: 10.2196/ijmr.2163] [Medline: 23612154]
9. Silva GS, Farrell S, Shandra E, Viswanathan A, Schwamm LH. The status of telestroke in the United States: a survey of currently active stroke telemedicine programs. Stroke 2012 Aug;43(8):2078-2085 [FREE Full text] [doi: 10.1161/STROKEAHA.111.645861] [Medline: 22700532]
10. Purvis RL, Sambamurthy V, Zmud RW. The assimilation of knowledge platforms in organizations: an empirical investigation. Organization Science 2001 Apr;12(2):117-135. [doi: 10.1287/orsc.12.2.117.10115]
11. 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]

12. Armstrong CP, Sambamurthy V. Information technology assimilation in firms: the influence of senior leadership and IT infrastructures. Information Systems Research 1999 Dec;10(4):304-327. [doi: 10.1287/isre.10.4.304]
13. Zhu K, Kraemer KL, Xu SX. The Process of innovation assimilation by firms in different countries: a technology diffusion perspective on e-business. Management Science 2006 Oct;52(10):1557-1576. [doi: 10.1287/mnsc.1050.0487]
14. Chatterjee D, Grewal R, Sambamurthy V. Shaping up for e-commerce: institutional enablers of the organizational assimilation of web technologies. MIS Quarterly 2002;26(2):65-89.
15. Hu PJH, Chau PYK, Sheng ORL. Adoption of telemedicine technology by health care organizations: an exploratory study. Journal of Organizational Computing and Electronic Commerce 2002 Sep;12(3):197-221. [doi: 10.1207/S15327744JOCE1203_01]
16. Furukawa MF, Raghu TS, Spaulding TJ, Vinze A. Adoption of health information technology for medication safety in U.S. Hospitals, 2006. Health Aff (Millwood) 2008;27(3):865-875 [FREE Full text] [doi: 10.1377/hlthaff.27.3.865] [Medline: 18474981]
17. Ford EW, Menachemi N, Phillips MT. Predicting the adoption of electronic health records by physicians: when will health care be paperless? J Am Med Inform Assoc 2006;13(1):106-112 [FREE Full text] [doi: 10.1197/jamia.M1913] [Medline: 16221936]
18. Davidson EJ, Chismar WG. The interaction of institutionally triggered and technology-triggered social structure change: an investigation of computerized physician order entry. MIS Quarterly 2007;31(4):739-758.
19. Ash JS, Bates DW. Factors and forces affecting EHR system adoption: report of a 2004 ACMI discussion. J Am Med Inform Assoc 2005;12(1):8-12 [FREE Full text] [doi: 10.1197/jamia.M1684] [Medline: 15492027]
20. Fichman RG, Kemerer CF. The illusory diffusion of innovation: an examination of assimilation gaps. Information Systems Research 1999 Sep;10(3):255-275. [doi: 10.1287/isre.10.3.255]
21. Meyer AD, Goes JB. Organizational assimilation of innovations: a multilevel contextual analysis. Academy of Management Journal 1988 Dec 01;31(4):897-923. [doi: 10.2307/256344]
22. Ash J. Organizational factors that influence information technology diffusion in academic health sciences centers. J Am Med Inform Assoc 1997;4(2):102-111 [FREE Full text] [Medline: 9067876]
23. Chau PYK, Hu PJH. Technology implementation for telemedicine programs. Communications of the ACM 2004;47(2):87-92.
24. Leonard KJ, Sittig DF. Improving information technology adoption and implementation through the identification of appropriate benefits: creating IMPROVE-IT. J Med Internet Res 2007;9(2):e9 [FREE Full text] [doi: 10.2196/jmir.9.2.e9] [Medline: 17513287]
25. Davidson EJ, Heslinga D. Bridging the IT adoption gap for small physician practices: an action research study on electronic health records. Information Systems Management 2006 Dec 22;24(1):15-28. [doi: 10.1080/10580530601036786]
26. Cho S, Mathiassen L. The role of industry infrastructure in telehealth innovations: a multi-level analysis of a telestroke program. Eur J Inf Syst 2007 Dec;16(6):738-750. [doi: 10.1057/palgrave.ejis.3000718]
27. Dion JE. Management of ischemic stroke in the next decade: stroke centers of excellence. J Vasc Interv Radiol 2004 Jan;15(1 Pt 2):S133-S141. [Medline: 15101522]
28. Schwamm LH, Pancioli A, Acker JE, Goldstein LB, Zorowitz RD, Shephard TJ, American Stroke Association's Task Force on the Development of Stroke Systems. Recommendations for the establishment of stroke systems of care: recommendations from the American Stroke Association's Task Force on the Development of Stroke Systems. Stroke 2005 Mar;36(3):690-703 [FREE Full text] [doi: 10.1161/01.STR.0000158165.42884.4F] [Medline: 15689577]
29. O'Toole LJ, Slade CP, Brewer GA, Gase LN. Barriers and facilitators to implementing primary stroke center policy in the United States: results from 4 case study states. Am J Public Health 2011 Mar;101(3):561-566. [doi: 10.2105/AJPH.2010.197954] [Medline: 21233430]
30. Rogove HJ, McArthur D, Demaerschalk BM, Vespa PM. Barriers to telemedicine: survey of current users in acute care units. Telemed J E Health 2012;18(1):48-53. [doi: 10.1089/tmj.2011.0071] [Medline: 22082107]
31. Eissa A, Krass I, Bajorek B. Barriers to the utilization of thrombolysis for acute ischaemic stroke. J Clin Pharm Ther 2012 Aug;37(4):399-409. [doi: 10.1111/j.1365-2710.2011.01329.x] [Medline: 22384796]
32. Kwan J, Hand P, Sandercock P. A systematic review of barriers to delivery of thrombolysis for acute stroke. Age Ageing 2004 Mar;33(2):116-121 [FREE Full text] [doi: 10.1093/ageing/afh064] [Medline: 14960425]
33. Miles MB, Huberman AM. Qualitative Data Analysis: An Expanded Sourcebook. Thousand Oaks: Sage Publications; 1994.
34. Yin RK. Case Study Research: Design and Methods. Thousand Oaks, CA: Sage Publications Inc; 2003.
35. Wang S, Lee SB, Pardue C, Ramsingh D, Waller J, Gross H, et al. Remote evaluation of acute ischemic stroke: reliability of National Institutes of Health stroke scale via telestroke. Stroke 2003 Sep 18;34(10):e188-e191 [FREE Full text] [doi: 10.1161/01.STR.0000091847.82140.9D]
36. Wang S, Gross H, Lee SB, Pardue C, Waller J, Nichols FT, et al. Remote evaluation of acute ischemic stroke in rural community hospitals in Georgia. Stroke 2004 Jul;35(7):1763-1768 [FREE Full text] [doi: 10.1161/01.STR.0000131858.63829.6e] [Medline: 15166386]
37. Hess DC, Wang S, Hamilton W, Lee S, Pardue C, Waller JL, et al. REACH. Stroke 2005;36(9):2018-2020.
38. Hess DC, Wang S, Gross H, Nichols FT, Hall CE, Adams RJ. Telestroke: extending stroke expertise into underserved areas. Lancet Neurol 2006 Mar;5(3):275-278. [doi: 10.1016/S1474-4422(06)70377-5] [Medline: 16488383]

39. Cho S, Khasanshina EV, Mathiassen L, Hess DC, Wang S, Stachura ME. An analysis of business issues in a telestroke project. J Telemed Telecare 2007;13(5):257-262. [doi: 10.1258/135763307781458930] [Medline: 17697514]
40. Edwards LL. Using tPA for acute stroke in a rural setting. Neurology 2007 Jan 23;68(4):292-294. [doi: 10.1212/01.wnl.0000253190.40728.0c] [Medline: 17242335]
41. Switzer JA, Hall C, Gross H, Waller JL, Nichols FT, Wang S, et al. A web-based telestroke system facilitates rapid treatment of acute ischemic stroke patients in rural emergency departments. J Emerg Med 2009 Jan;36(1):12-18. [doi: 10.1016/j.jemermed.2007.06.041] [Medline: 18242925]
42. Cho S, Mathiassen L, Robey D. Dialectics of resilience: a multi-level analysis of a telehealth innovation. J Inf Technol 2006 Dec 19;22(1):24-35. [doi: 10.1057/palgrave.jit.2000088]
43. Cho S, Mathiassen L, Gallivan M. Crossing the diffusion chasm: from invention to penetration of a telehealth innovation. Information Technology & People 2009;22(4):351-366. [doi: 10.1108/09593840911002450]
44. Switzer JA, Akinwuntan A, Waller JL, Nichols FT, Hess DC, Bruno A. Impact of primary stroke center certification on location of acute ischemic stroke care in Georgia. Stroke 2012 May;43(5):1415-1417 [FREE Full text] [doi: 10.1161/STROKEAHA.111.648378] [Medline: 22363063]
45. Lazaridis C, DeSantis SM, Jauch EC, Adams RJ. Telestroke in South Carolina. J Stroke Cerebrovasc Dis 2013 Oct;22(7):946-950. [doi: 10.1016/j.jstrokecerebrovasdis.2011.11.008] [Medline: 22196873]
46. Adams RJ, Debenham E, Chalela J, Chimowitz M, Hays A, Hill C, et al. REACH MUSC: a telemedicine facilitated network for stroke: initial operational experience. Front Neurol 2012;3:33 [FREE Full text] [doi: 10.3389/fneur.2012.00033] [Medline: 22435064]
47. Kazley AS, Wilkerson RC, Jauch E, Adams RJ. Access to expert stroke care with telemedicine: REACH MUSC. Front Neurol 2012;3:44 [FREE Full text] [doi: 10.3389/fneur.2012.00044] [Medline: 22461780]
48. Alberts MJ. Recommendations for the establishment of primary stroke centers. JAMA 2000 Jun 21;283(23):3102-3109. [doi: 10.1001/jama.283.23.3102]
49. Jauch EC, Saver JL, Adams HP, Bruno A, Connors J, Demaerschalk BM, American Heart Association Stroke Council, Council on Cardiovascular Nursing, Council on Peripheral Vascular Disease, Council on Clinical Cardiology. Guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2013 Mar;44(3):870-947 [FREE Full text] [doi: 10.1161/STR.0b013e318284056a] [Medline: 23370205]
50. Gogan JL, Garfield M. Deploying resources to save lives: partners tele-stroke service. 2011 Presented at: European Conference on Information Systems (ECIS); June 9-11, 2011; Helsinki, Finland p. 1-12.
51. de Bustos EM, Moulin T, Audebert HJ. Barriers, legal issues, limitations and ongoing questions in telemedicine applied to stroke. Cerebrovasc Dis 2009;27 Suppl 4:36-39. [doi: 10.1159/000213057] [Medline: 19546540]
52. Meurer W, Majersik J, Frederiksen S, Kade A, Sandretto A, Scott P. Provider perceptions of barriers to the emergency use of tPA for acute ischemic stroke: a qualitative study. BMC Emerg Med 2011;11:5 [FREE Full text] [doi: 10.1186/1471-227X-11-5] [Medline: 21548943]
53. Mason J. Qualitative Researching. Second Edition. Thousand Oaks, CA: Sage Publications Inc; 2002.
54. Lincoln YS, Guba EG. Naturalistic Inquiry. First Edition. Newbury Park, CA: Sage Publications Inc; 1985.
55. Devers KJ. How will we know "good" qualitative research when we see it? Beginning the dialogue in health services research. Health Serv Res 1999 Dec;34(5 Pt 2):1153-1188 [FREE Full text] [Medline: 10591278]
56. Paulk MC, Curtis B, Chrissis MB, Weber CV. Capability maturity model, version 1.1. IEEE Softw 1993 Jul;10(4):18-27. [doi: 10.1109/52.219617]

Abbreviations

CT: computed tomography
ED: emergency department
EMS: emergency medical services
GRU: Georgia Regents University
ICU: intensive care unit
IT: information technology
MUSC: Medical University of South Carolina
REACH: Remote Evaluation of Acute isCHemic stroke
tPA: tissue plasminogen activator

Edited by G Eysenbach; submitted 15.10.13; peer-reviewed by L Wechsler, C Parra, D Jacobs; comments to author 05.11.13; revised version received 31.12.13; accepted 10.01.14; published 15.04.14.

Please cite as:
Singh R, Mathiassen L, Switzer JA, Adams RJ
Assimilation of Web-Based Urgent Stroke Evaluation: A Qualitative Study of Two Networks
JMIR Med Inform 2014;2(1):e6
URL: http://medinform.jmir.org/2014/1/e6/
doi:10.2196/medinform.3028
PMID:25601232

Viewpoint

A Software System to Collect Expert Relevance Ratings of Medical Record Items for Specific Clinical Tasks

H Benjamin Harvey[1,2], MD, JD; Arun Krishnaraj[3], MD, MPH; Tarik K Alkasab[1,2], MD, PhD

[1]Department of Radiology, Massachusetts General Hospital, Boston, MA, United States

[2]Harvard Medical School, Boston, MA, United States

[3]Department of Radiology, University of Virginia, Charlottesville, VA, United States

Corresponding Author:
H Benjamin Harvey, MD, JD
Department of Radiology
Massachusetts General Hospital
55 Fruit Street, FND 216
Boston, MA, 02114
United States
Phone: 1 617 724 4255
Fax: 1 617 726 3077
Email: hbharvey@partners.org

Abstract

Development of task-specific electronic medical record (EMR) searches and user interfaces has the potential to improve the efficiency and safety of health care while curbing rising costs. The development of such tools must be data-driven and guided by a strong understanding of practitioner information requirements with respect to specific clinical tasks or scenarios. To acquire this important data, this paper describes a model by which expert practitioners are leveraged to identify which components of the medical record are most relevant to a specific clinical task. We also describe the computer system that was created to efficiently implement this model of data gathering. The system extracts medical record data from the EMR of patients matching a given clinical scenario, de-identifies the data, breaks the data up into separate medical record items (eg, radiology reports, operative notes, laboratory results, etc), presents each individual medical record item to experts under the hypothetical of the given clinical scenario, and records the experts' ratings regarding the relevance of each medical record item to that specific clinical scenario or task. After an iterative process of data collection, these expert relevance ratings can then be pooled and used to design point-of-care EMR searches and user interfaces tailored to the task-specific needs of practitioners.

(JMIR Med Inform 2014;2(1):e3) doi:10.2196/medinform.3204

KEYWORDS

medical informatics; health information management; computerized medical records system

Introduction

Adoption of electronic medical records (EMR) has increased dramatically over the past decade, driven in part by sizeable federal subsidies [1,2]. This growth has meant an attendant dramatic increase in the amount and variability of patient data stored in a typical patient's EMR, creating difficult challenges related to data organization and presentation. As a result, the necessary information to answer a clinical question may be spread among several potentially unstructured documents, requiring a practitioner to undergo a laborious EMR search process. This, in turn, can decrease efficiency, increase medical errors, and generate dissatisfaction among practitioners, potentially negating the safety and efficiency improvements associated with EMR use [3-6]. Difficult-to-navigate EMRs may also contribute to the problem of rising health care costs, because practitioners who are unaware of information contained within the EMR may be more likely to order unnecessary or duplicate tests and procedures [7].

In light of these challenges, the efficiency and accuracy of practitioner data retrieval should be a key focus in the ongoing design of clinical EMR systems and supporting software tools. The addition of advanced EMR search capabilities, such as keyword searches, have improved radiologist efficiency and have the potential to improve patient outcomes [8,9]. To have even greater impacts on clinical care and to improve value, the next generation of EMR technology needs to go beyond keyword searchability and instead present practitioners with a filtered

view of the medical record that is germane to their task-specific clinical needs. For example, a radiologist interpreting a magnetic resonance image of a patient's liver will be interested in a subset of the medical record focused on hepatic and other abdominal issues, along with any history of malignancy. However, a neurologist seeing the same patient for the management of Parkinson's disease will be interested in a different set of notes, reports, and data. Ideal EMR search algorithms and user interfaces should differentiate between the two practitioners and clinical scenarios. Multiple groups have kick-started this process by developing and validating automated EMR search strategies and data displays for specific clinical tasks, including identification of preprocedural and preoperative risk factors for complications, prediction of long-term mortality of patients admitted to the hospital, and the treatment of intensive care unit patients and neuro-oncology patients [10-14].

A major challenge in designing these task-specific EMR tools for clinical use is obtaining the information about which components of the EMR are most relevant to practitioners in specific clinical scenarios. To overcome this, we propose a strategy for collecting these relevancy data. The proposed approach starts by extracting and de-identifying medical record data from an actual patient in a given clinical scenario. The medical record is then disassembled into component medical record items (eg, radiology reports, operative notes, laboratory results, etc), which are individually presented to a panel of clinical experts. Each medical record item is rated by the experts for its relevance to a specific clinical scenario or task. This process is performed iteratively for multiple patients in the same clinical scenario, thereby creating a robust body of expert-provided relevancy data that indicates which medical record items are most valuable in that particular clinical scenario. The expert-generated relevance data can then be used to design and validate EMR search algorithms and user interfaces tailored to that clinical scenario.

In this study, we describe a software system that we created to implement this process of data collection. We hope this work will serve as the basis for ongoing efforts to improve the value of EMR technology for patients and practitioners.

Methods

Tool to Extract Electronic Medical Record Data

A tool was created to extract, de-identify, and format data from our institution's EMR system according to the defined schema. The first version of this tool was designed around the clinical scenario of a radiologist interpreting an abdominal computed tomography (CT) scan for a patient with a clinical history of abdominal pain. When an index abdominal CT scan matching this specific clinical scenario was identified, the queriable patient inference dossier EMR search/aggregation tool was used to find and extract all radiology reports, operative notes, laboratory results, pathology reports, endoscopy reports, and microbiology results for the given patient within a period extending from 2 years prior to the index radiology exam to 2 years after the index exam [8,9]. These items were selected because they are separately identifiable in our institution's medical record system and were thought likely to be relevant to common subspecialty clinical situations. A universally unique identifier (UUID) was assigned to the scenario as a whole and for each individual medical record item [15]. The tool automatically removed identifying patient information including the patient's first and last names, any dates, all physician names, and all identifying numbers (eg, medical record numbers, accession numbers, phone/fax numbers, zip codes, etc). Patient demographic data were reduced to sex and age, with ages greater than or equal to 90 years reported as 89 years to reduce identifiability. Because no look-up table was maintained, re-identification of the patient record was not possible. Although this may reduce opportunities to add additional information to a specific scenario later, it was judged that protecting patient privacy outweighed this loss. The resulting structure was written to an XML file of the format specified in the scenario schema.

Tool to Collect Rater Scores

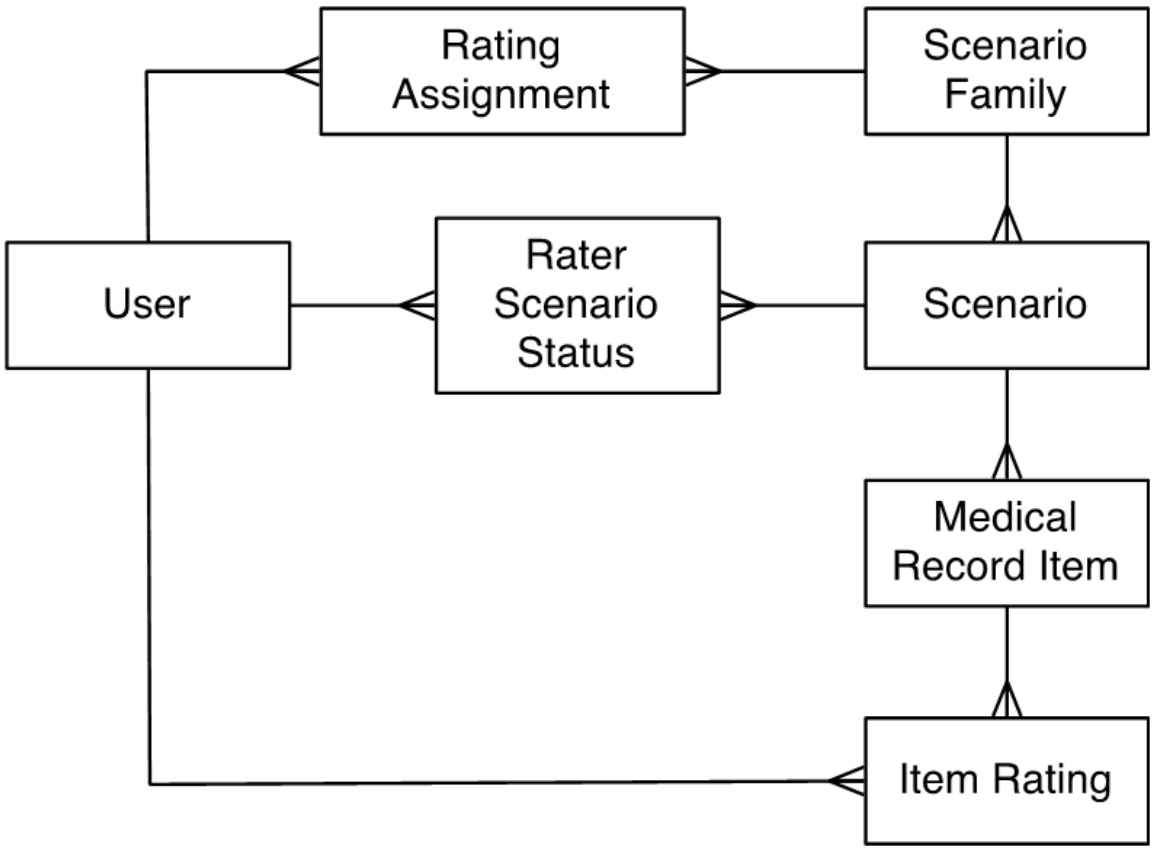
A separate tool was created to manage the collection of expert ratings. The tool was designed to represent incoming sample medical record data sets, information on raters, and the assigned rating scores. The data model to represent these data is presented in Figure 1. This model allows internal representation of incoming medical record data sets as defined in the scenario schema, and exporting of the data into a file according to the scenario family ratings schema. The data model was centered on the ScenarioFamily: data structure; that is, groups of different patients' de-identified medical records selected and extracted based on a shared clinical context. Each individual case/patient was represented by a Scenario data structure, which in turn is made up of the individual EMR entries for that patient, the MedicalRecordItem objects. When an expert registers to be a rater, a User object was created. Users were then assigned to rate ScenarioFamily objects; this connection was a RatingAssignment. The user's progress toward completing the Scenarios in the task list of assignments was tracked by RaterScenarioStatus objects. The actual relevance ratings were stored as ItemRating objects. ScenarioFamily objects were assigned UUIDs as needed, as were Users.

Three system interfaces were necessary to permit rater and administrative interaction with the system. The first interface, or task list, showed raters their current set of assignments so they could advance to the next task when finished. The central interface of the system laid out the clinical context for the rater and asked them to assign a relevance score to a medical record item. Finally, an administrative interface was needed for manipulating the scenario families and user assignments and monitoring progress of tasks across the system.

We used open source technologies to implement this system, specifically choosing the Ruby on Rails Web application framework backed by the SQLite3 database engine [16,17]. This resulted in a model-view-controller architecture that could be easily understood and implemented with a variety of platforms and frameworks. The open-source database means that many analysis tools and other software programs can access the generated data as needed. All pages complied with the HTML5 standard to ensure optimal compatibility with modern Web browsers. The Twitter Bootstrap front-end framework and jQuery JavaScript library enabled creation of a richer front end

[18]. A Devise authentication plugin was used to manage the process of creating and authenticating raters [19]. We deployed our implementation on a Mac OS X system, but it is possible to deploy the system on most Unix/Linux-based operating systems.

Figure 1. Rating collection tool data model. Inside the collection tool, each set of sample medical records is represented by a ScenarioFamily object, which contains many Scenario objects, which in turn contains many MedicalRecordItems. Expert raters are represented by User objects and are associated with a ScenarioFamily via a RatingAssignment. Raters create an ItemRating object for each MedicalRecordItem within Scenarios belonging to each ScenarioFamily to which they are assigned.



Results

After obtaining institutional review board approval, the utility of the data extraction tool was demonstrated by extracting data from our institution's EMR and successfully generating de-identified sample medical record data sets. In general, the automated redaction process was quite effective. However, to ensure maximal protection of potentially sensitive patient information, each data set was further manually examined for residual protected health information, which was then redacted by hand. De-identified data sets were then imported into the database of the rating collection tool using a separate command line utility. The rating collection tool was not actively connected to the EMR system. The scenario data sets were grouped into scenario families; each scenario is composed of a specific example of the clinical context defined by the scenario family.

Persons serving as expert reviewers were directed to the system home page, where they could register for a rater account by providing basic personal information to a Web interface. An administrator then assigned each registered user to scenario families based on their clinical expertise using a command line tool. Once scenario assignments were made, the user logged into the system and could see their list of scenario assignments (Figure 2). The rater chose which scenario to work on by clicking the appropriate item from the list.

After choosing a scenario from the task list, the system brings the user into the main rating interface (Figure 3). The rater is shown the name of the scenario family/clinical context being considered. The central "Context" column shows the specific clinical context of the sample patient whose medical record items the expert user must rate. For example, for the clinical scenario family of a radiologist interpreting abdominal CT scans performed for a clinical history of abdominal pain, the central "Context" column would be the respective clinical history (and possibly the report) of a specific abdominal CT scan that the rater should envision wanting to interpret in that clinical context. Along the left column is the list of medical record items extracted from the EMR of the sample patient in that clinical context. Within the list are both the medical record items that still need to be rated, along with the items that have already been assigned a relevance score by the rater. The rater can scroll up and down this list to review the items that they have already rated and the relevance scores assigned to those items. An

indicator of the rater's progress through a given clinical scenario (ie, an individual patient) and the scenario family more broadly (ie, the assigned cohort of the sample patients) is shown in the upper right portion of the screen.

When the rater clicks on a specific medical record item in the list on the left-hand column of the screen, the system presents the rater with the medical record item in full detail in the right-hand column along with the choices for rating relevance. The rating relevancy choices are represented both as words ("irrelevant," "unlikely relevant," "probably relevant," and "certainly relevant") and as a number of filled-in stars (0-3). Once the rater clicks on a relevancy rating, the system automatically presents the next medical record item in the right-hand column. When all of the medical record items have been rated in a given scenario, the rater is taken to the first item in the next scenario in the scenario family. Likewise, when all of the scenarios in the scenario family have been rated, the rater is returned to the home page, where they can choose to proceed to the next uncompleted assignment.

An administrative user can track the progress of raters through their assigned scenario families via an overview interface (Figure 4). This interface displays a list of the scenario families known to the system and abbreviated UUIDs for its component scenarios. The overview interface displays the current progress of each rater in their task list, for all raters assigned to a particular scenario family. For each family, links are provided to add a new rater, upload a new scenario, or download the current ratings data.

Finally, once the assigned panel of raters has rated the relevant items for the individual scenarios in a scenario family, an XML file containing the relevant rating data can be extracted. These ratings data can then be used to design and validate tailored medical record search strategies and user interfaces for a given clinical task.

Figure 2. Post log-in page showing assigned tasks. After a user logs into the system, they are presented with a list of their currently assigned clinical scenarios. They can use these links to move directly to the rating interface.

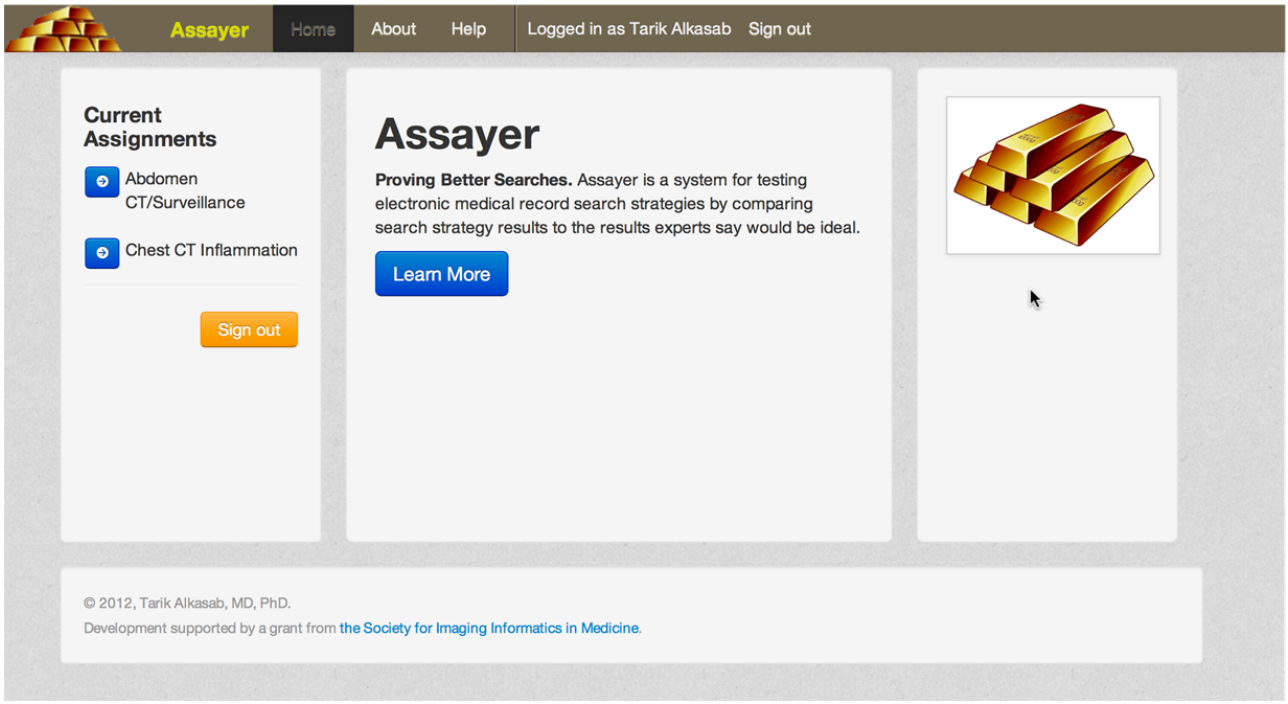


Figure 3. Expert rating interface. An expert rater considers a particular scenario family and a specific clinical context and assesses the relevance of the items in the medical record to the given scenario. Relevance is rated on a 4-step scale: irrelevant, unlikely relevant, probably relevant, and certainly relevant.

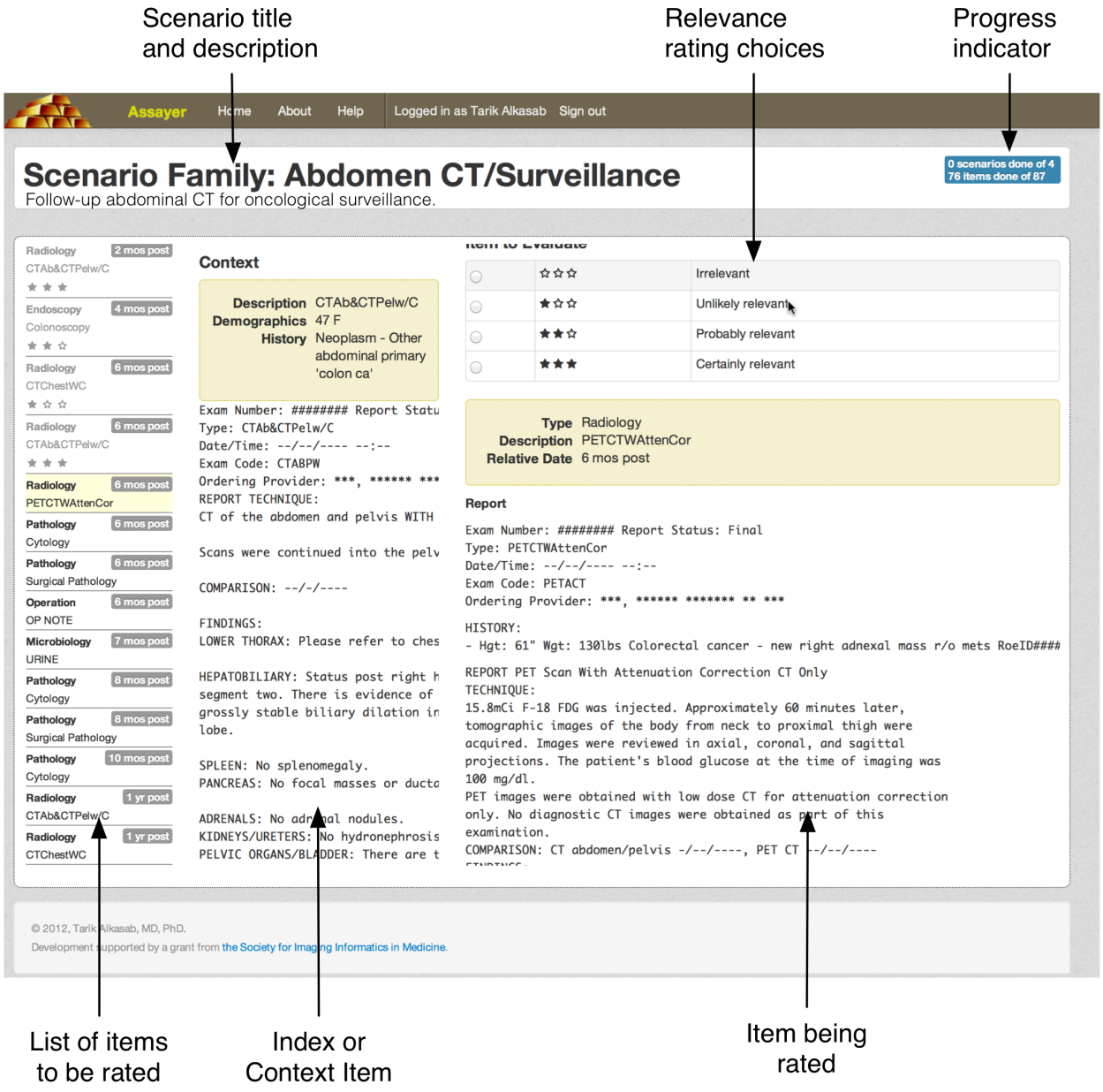
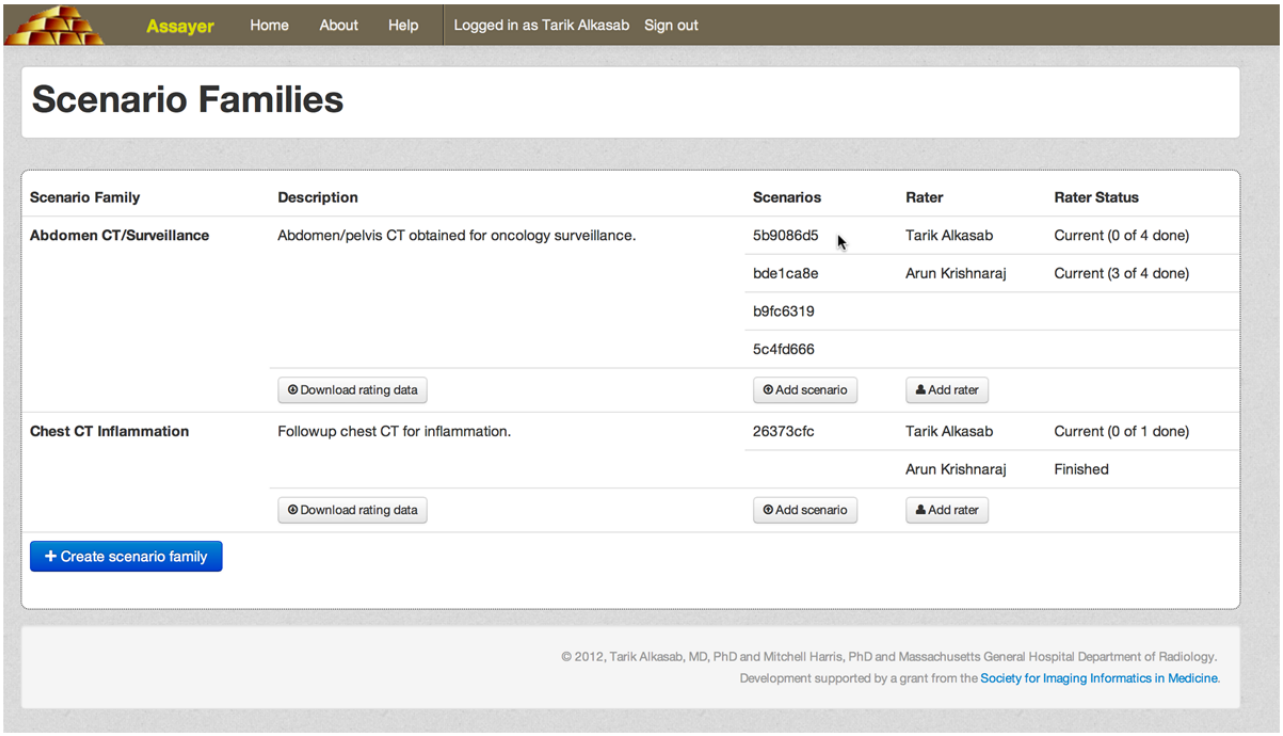


Figure 4. Administrator monitoring interface. An administrator keeping track of activity on the service can view the monitoring page, which shows all of the active scenario families, the scenarios making them up, the raters assigned to evaluate the items, and their progress in rating all of the items. Tools for defining new scenario families, uploading scenarios, and adding raters are also available from this page.



Discussion

Principal Results

We have successfully designed and implemented a system for extracting exemplar medical record items, such as laboratory results and operative reports, from the EMR and obtaining task-specific expert relevance ratings for those items. The extraction tool pulls a subset of the medical records for a patient, de-identifies it, and formats it so that it can be included in the rating collection tool. The rating collection tool then allows a clinical expert, such as a radiologist for the clinical application described herein, to review and rate the component parts of even complex medical records, thereby highlighting the items in the medical record that are most relevant to a specific clinical task. Both the de-identified clinical information and the expert-supplied relevance ratings are captured, organized, and exported in a format that can be used for search strategy optimization and the design of tailored EMR user interfaces.

The design of this system was based on a few guiding principles. First, the system should be built using easily available, well-understood open source tools according to standard design patterns. Second, the system should easily fit into a broader framework for designing medical record searches and user interfaces, including, but not limited to, importing and exporting open formats to simplify data interchange. Finally, the system's interface with expert raters should be simple and efficient.

Based on these principles, we created a system to reduce the effort associated with the collection of expert ratings data, while ensuring the accuracy and robustness of the data collected. Recognizing the high value of an expert's time, efficiency of the process was an absolute requirement. Thus, whenever an expert begins a session, he or she is moved into the process of examining and rating medical record items as rapidly and efficiently as possible. The tool also allows multiple experts to evaluate the same body of de-identified patient data. Having multiple raters review the same medical record items reduces the effect of individual rater idiosyncrasies. Moreover, new scenarios can easily be added to a family to reduce the effect of specific variations within a single source medical record, if needed. Last of all, the system allowed experts to be matched with clinical scenarios specific to their expertise.

Limitations

With the software system implemented, the most important challenges to actually putting the system into use revolve around selecting an appropriate clinical scenario and recruiting appropriate experts. With regard to clinical scenarios, it is important to make the scenario specific enough so that the expert raters think they are making a concrete decision about relevance rather than an abstract one. Expert recruitment, on the other hand, benefits from selecting both highly specialized experts and more generalist practitioners. To maintain the raters' interest, everything possible should be done to reduce administrative overhead on the raters.

Conclusions

Looking forward, it will be important to develop a capacity to facilitate interchange between different sites or installations to maximize the generalizability of the collected ratings data. Even though the medical record items are de-identified, such interinstitutional collaborations would likely require access to be restricted to a predefined group rather than the world at large due to the potentially sensitive nature of the information. This could be accomplished by allowing each site to maintain a list

of partner sites where a combination of sample medical record items and ratings data could be securely exchanged. Once established, exporting scenario data, scenario family data, and ratings data could then be available via a website. In this way, a broader range of experts and sample clinical items could be assessed, leading to a more robust body of expert ratings data.

We envision a multicenter project to collect expert relevance ratings for several clinical scenarios common to radiology. In this project, sample medical record items will be pulled from each center and pooled into a common sample set for each clinical scenario. The pooled expert relevancy ratings data can then be used to validate candidate search strategies and eventually to develop filtered views of the medical record specific for clinical tasks commonly faced by radiologists. Obviously, such interinstitution synergy poses many challenges, not least of which is the incompatibility of medical record data formats. As institutions begin such collaborations, it is important to define standards for representation of the medical record information. This could even be expanded to account for international differences. It will also likely be necessary to expand the data model to clearly state which experts should be allowed to view data from which partner institutions. These additional efforts would be rewarded by a much richer set of sample medical record data and ratings.

The approach we outlined emphasizes a “wisdom of crowds” data-driven approach to identifying likely-to-be relevant medical record information rather than an expert-driven methodology. We designed the tool described herein to gather this collective intelligence because it is much harder to incorporate such information into a design process. In fact, we believe that the most effective search strategies will be generated by starting with hyper-local experts, who define relevance based on their specialized experience, and then proving and testing their designs against crowd-sourced data. Far from denigrating the potential contribution of individual innovation, we hope to provide a way to hone those contributions. Future versions of the software could allow the raters to provide specific comments and notes to search designers to further spur these efforts.

In sum, context-specific EMR searches and user interfaces have the potential to increase the efficiency and safety and reduce the cost of health care delivery. To achieve these ends, development of these tools must be data-driven and influenced by an understanding of practitioner information requirements. The data collected using the herein described software tool can serve as the basis for acquiring this essential guidance, with the ultimate goal of creating tools that allow physicians to rapidly and effectively navigate EMR systems. By providing this as an open-source tool with open formats for data interchange, we hope to bolster the adoption of the process through interinstitutional synergy.

Conflicts of Interest

None declared.

References

1. Hsiao C, Hing E, Socey C, Cai B. National Center for Health Statistics (NCHS). Electronic Health Record Systems and Intent to Apply for Meaningful Use Incentives Among Office-based Physician Practices: United States, 2001-2011 URL: http://www.cdc.gov/nchs/data/databriefs/DB79.pdf [accessed 2013-12-21] [WebCite Cache ID 6M2C4RDAe]
2. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010 Aug 5;363(6):501-504. [doi: 10.1056/NEJMp1006114] [Medline: 20647183]
3. Ahmed A, Chandra S, Herasevich V, Gajic O, Pickering BW. The effect of two different electronic health record user interfaces on intensive care provider task load, errors of cognition, and performance. Crit Care Med 2011 Jul;39(7):1626-1634. [doi: 10.1097/CCM.0b013e31821858a0] [Medline: 21478739]
4. Boonn WW, Langlotz CP. Radiologist use of and perceived need for patient data access. J Digit Imaging 2009 Aug;22(4):357-362 [FREE Full text] [doi: 10.1007/s10278-008-9115-2] [Medline: 18459002]
5. Bates DW, Gawande AA. Improving safety with information technology. N Engl J Med 2003 Jun 19;348(25):2526-2534. [doi: 10.1056/NEJMsa020847] [Medline: 12815139]
6. Zlabek JA, Wickus JW, Mathiason MA. Early cost and safety benefits of an inpatient electronic health record. J Am Med Inform Assoc 2011;18(2):169-172 [FREE Full text] [doi: 10.1136/jamia.2010.007229] [Medline: 21292703]
7. Arenson RL. Recommendations for additional imaging in radiology reports: radiologists' self-referral or good clinical practice? Radiology 2009 Nov;253(2):291-292. [doi: 10.1148/radiol.2532091216] [Medline: 19864523]
8. Zalis M, Harris M. Advanced search of the electronic medical record: augmenting safety and efficiency in radiology. J Am Coll Radiol 2010 Aug;7(8):625-633. [doi: 10.1016/j.jacr.2010.03.011] [Medline: 20678732]
9. Seyfried L, Hanauer DA, Nease D, Albeiruti R, Kavanagh J, Kales HC. Enhanced identification of eligibility for depression research using an electronic medical record search engine. Int J Med Inform 2009 Dec;78(12):e13-e18 [FREE Full text] [doi: 10.1016/j.ijmedinf.2009.05.002] [Medline: 19560962]
10. Campbell EJ, Krishnaraj A, Harris M, Saini S, Richter JM. Automated before-procedure electronic health record screening to assess appropriateness for GI endoscopy and sedation. Gastrointest Endosc 2012 Oct;76(4):786-792. [doi: 10.1016/j.gie.2012.06.003] [Medline: 22901989]

11. Alsara A, Warner DO, Li G, Herasevich V, Gajic O, Kor DJ. Derivation and validation of automated electronic search strategies to identify pertinent risk factors for postoperative acute lung injury. Mayo Clin Proc 2011 May;86(5):382-388 [FREE Full text] [doi: 10.4065/mcp.2010.0802] [Medline: 21531881]
12. Singh B, Singh A, Ahmed A, Wilson GA, Pickering BW, Herasevich V, et al. Derivation and validation of automated electronic search strategies to extract Charlson comorbidities from electronic medical records. Mayo Clin Proc 2012 Sep;87(9):817-824 [FREE Full text] [doi: 10.1016/j.mayocp.2012.04.015] [Medline: 22958988]
13. Ahmed A, Chandra S, Herasevich V, Gajic O, Pickering BW. The effect of two different electronic health record user interfaces on intensive care provider task load, errors of cognition, and performance. Crit Care Med 2011 Jul;39(7):1626-1634. [doi: 10.1097/CCM.0b013e31821858a0] [Medline: 21478739]
14. Hsu W, Taira RK, El-Saden S, Kangarloo H, Bui AA. Context-based electronic health record: toward patient specific healthcare. IEEE Trans Inf Technol Biomed 2012 Mar;16(2):228-234. [doi: 10.1109/TITB.2012.2186149] [Medline: 22395637]
15. Network Working Group. A Universally Unique IDentifier (UUID) URN Namespace.: The Internet Society; 2005. URL: http://tools.ietf.org/html/rfc4122 [accessed 2013-12-21] [WebCite Cache ID 6M2CFBN2z]
16. Heinemeier Hansson D. Ruby on Rails. 2003. URL: http://rubyonrails.org/ [accessed 2013-12-21] [WebCite Cache ID 6M2CT3cDm]
17. SQLite, Version 3.8.3.1. URL: http://sqlite.org/ [accessed 2013-12-21] [WebCite Cache ID 6M2CZQLsA]
18. jQuery Foundation. jQuery. URL: http://jquery.com/ [accessed 2013-12-21] [WebCite Cache ID 6M2Cnq5ee]
19. Plataformatec. Devise. URL: https://github.com/plataformatec/devise [accessed 2013-12-21] [WebCite Cache ID 6M2Cuz4xh]

Abbreviations

CT: computed tomography
EMR: electronic medical record
UUID: universally unique identifier

Edited by G Eysenbach; submitted 21.12.13; peer-reviewed by M Wong, YM Tan; comments to author 12.01.14; accepted 28.01.14; published 28.02.14.

Please cite as:
Harvey HB, Krishnaraj A, Alkasab TK
A Software System to Collect Expert Relevance Ratings of Medical Record Items for Specific Clinical Tasks
JMIR Med Inform 2014;2(1):e3
URL: http://medinform.jmir.org/2014/1/e3/
doi:10.2196/medinform.3204
PMID:25600925

Viewpoint

Use of Expert Relevancy Ratings to Validate Task-Specific Search Strategies for Electronic Medical Records

Harlan Harvey[1], MD, JD; Arun Krishnaraj[2], MD, MPH; Tarik K Alkasab[1], MD, PhD

[1]Department of Radiology, Massachusetts General Hospital, Boston, MA, United States

[2]Department of Radiology, University of Virginia, Charlottesville, VA, United States

Corresponding Author:
Harlan Harvey, MD, JD
Department of Radiology
Massachusetts General Hospital
55 Fruit Street, FND 216
Boston, MA, 02116
United States
Phone: 1 617 724 4255
Fax: 1 617 726 3077
Email: hbharvey@partners.org

Abstract

As electronic medical records (EMRs) grow in size and complexity, there is increasing need for automated EMR tools that highlight the medical record items most germane to a practitioner's task-specific needs. The development of such tools would be aided by gold standards of information relevance for a series of different clinical scenarios. We have previously proposed a process in which exemplar medical record data are extracted from actual patients' EMRs, anonymized, and presented to clinical experts, who then score each medical record item for its relevance to a specific clinical scenario. In this paper, we present how that body of expert relevancy data can be used to create a test framework to validate new EMR search strategies.

(JMIR Med Inform 2014;2(1):e4) doi:10.2196/medinform.3205

KEYWORDS

medical informatics; medical records systems; computerized; health information management

Introduction

Electronic Medical Records

As electronic medical records (EMR) become more common throughout the medical community, a wider variety of structured and unstructured data are being incorporated into them. Increasing EMR content has meant that some data necessary for clinical decision making are spread among several documents and repositories. This has the potential to increase practitioner workload, predispose to medical errors, and result in unnecessary utilization of health care resources [1,2]. In an attempt to reclaim efficiency, practitioners may lean on unreliable heuristics to obtain the answers they need.

Task-Specific Algorithms

Task-specific EMR search algorithms could ameliorate this situation by better addressing the diverse needs of practitioners [3-5]. However, one challenge in designing task-specific algorithms is finding a way to validate proposed search strategies prior to clinical implementation, given a lack of task-specific gold standards [6]. We have previously described a process for collecting context-specific expert relevancy ratings of medical record items [7]. The process relies on anonymized items of medical record data extracted from actual patients' EMRs, which are then presented to and rated by clinical experts based on the medical record items relevancy to a specific clinical task or scenario. The resultant relevancy data collected by the process can serve as the gold standard against which to evaluate search algorithms.

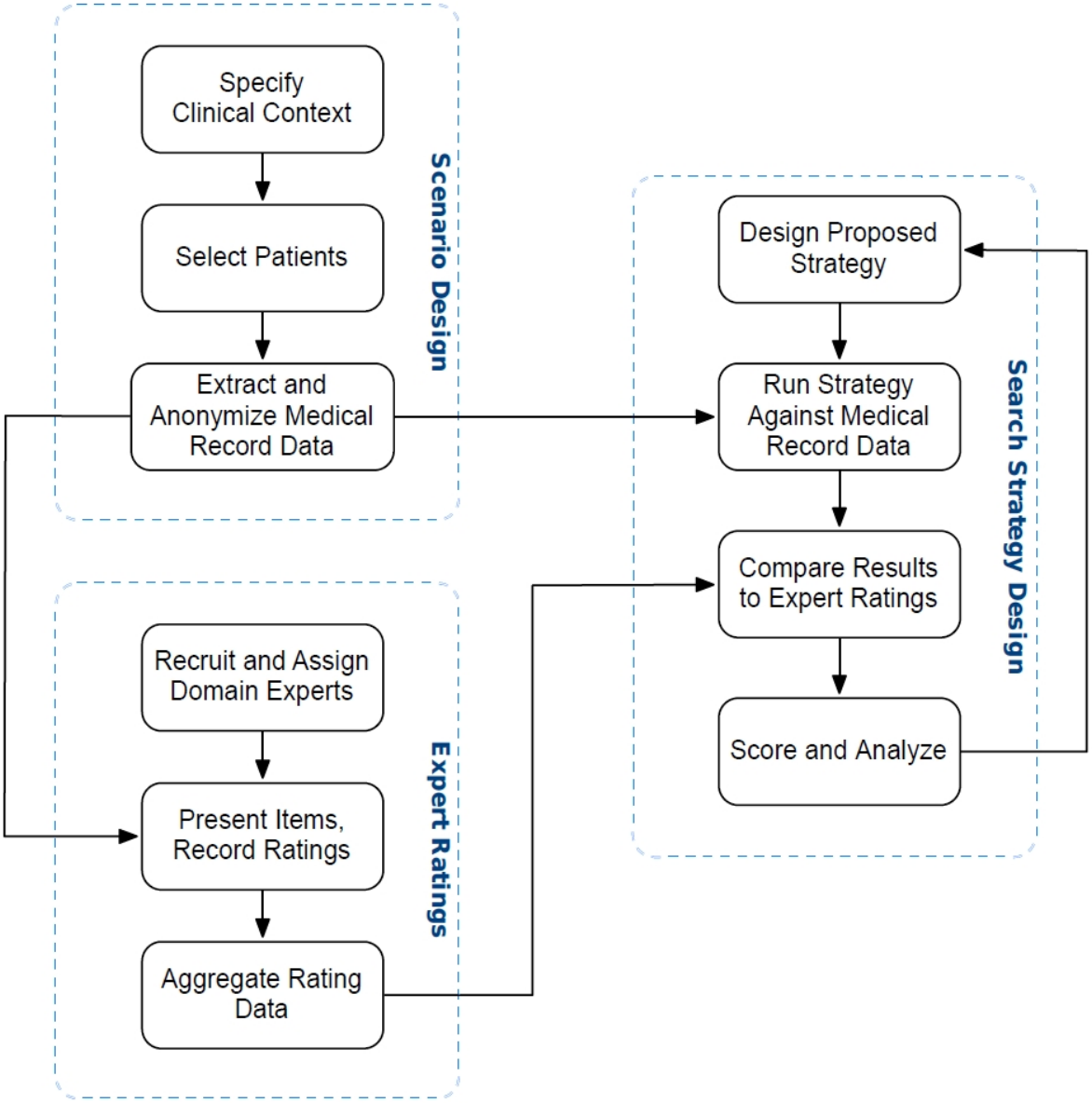
In this paper, we describe how the expert relevancy ratings data can be employed as a test framework to validate search strategies. We include proposed formats for transmitting data between separate steps and a preliminary algorithm for assessing the concordance between the "hits" from a search strategy and the expert relevance ratings.

Methods

The Three Main Subprocesses

There are three main subprocesses that are required to implement this vision for any given clinical scenario (Figure 1 shows these subprocesses). First, representative medical record data must be extracted from the EMR to serve as a target dataset. This dataset should incorporate a range of medical record items, chosen broadly from all items that might be available to a practitioner in the given clinical scenario. Second, this set of medical record items must be presented to a panel of clinical experts, who will rate each item for relevance in the given clinical scenario. Typically, these experts are clinical physicians who have been recruited because they frequently encounter the given clinical situation; their collected relevance ratings serve as a gold standard for the relevance of medical record items, which might or might not be included as a search result. Finally, a proposed search strategy will be generated and tested against the set of medical record items. The agreement between the items highlighted by the search algorithm and those rated as relevant by the experts can be computed and used as a performance metric for the search algorithm.

Figure 1. The flow of data through a process of validated medical record searches for a specific clinical context. For a defined clinical context, a set of representative patients is selected and medical record items are extracted and anonymized. These datasets are then presented to a panel of domain experts who generate a set of rating data. Meanwhile, an automated search to highlight relevant items is designed and then run against all of the anonymized medical record data to determine which items would be considered "hits." This result set is then compared with the expert relevance ratings and a normalized score is generated which quantifies the level of agreement between the search and the experts, which can then be used to design improvements in the search.



Example Medical Record Data

For a given clinical scenario, a set of matching patients can be selected. A sample of matching medical record items can then be extracted from the EMR system and anonymized. This set of medical record items for one patient is deemed a scenario, and can be expressed as an eXtensible Markup Language (XML) data file matching the following RELAX NG Compact open source schema found at the referenced link [8,9]. This defines a <clinical_scenario> as including patient demographic information and data regarding an index examination. Then, a list of <medical_record_item>s is listed, along with information about the type of record and the number of days between the record and the index exam.

A set of such scenarios that are examples of a single clinical scenario is termed a <scenario_family>, as defined by the open source schema found at the referenced link [10]. A <scenario_family> contains numerous <scenario_reference>s. This definition, together with the relevant defined scenarios, forms the dataset against which a search strategy can be run.

Expert Rating Data

Once a set of medical record data is available, it can be presented to a group of experts. The expert panel is made up of clinicians from the particular medical specialty tasked with the clinical scenario of interest. For example, if the method were being employed to identify medical record items pertinent to the clinical task of interpreting an MRI examination of the liver, the expert panel would be made up of abdominal radiologists knowledgeable in the clinical information germane to that task. The experts will rate each item for its relevance to the particular scenario along a four-step scale. The steps are labeled-"Irrelevant," "Unlikely relevant," "Probably relevant," and "Certainly relevant." These rating data can be gathered into an XML file that matches the open source schema, found at the referenced link [11]. The data are stored as a hierarchy of <scenario_family_ratings>, <scenario_ratings>, <rater_data>, and individual <item_rating>s.

Search Strategy Results

The results of running a given search strategy against the medical record items contained in a <scenario_family> can be represented using an open source schema found at the referenced link [12]. The schema organizes a series of <item_result> elements, each of which indicates whether the given search strategy would include the given item as a hit or not.

Strategy Scoring Metric

A scoring metric was developed for describing the extent of agreement between results returned by a particular search strategy and the expert rating data. The strategy is based on calculation of the kappa statistic [13]. This measure will be highest when experts agree on the relevance of an item of medical record data and the search strategy appropriately includes or excludes the item. Specifically, the statistic will increase monotonically with increasing agreement between a tested search strategy and the expert raters. After expert rating data have been collected for a given test set of medical record items and a candidate search strategy is tested against that same test set, these metrics are calculated to assess the performance of the candidate strategy.

Results

The overall performance of the search strategy is captured by a single metric, S_{total}. Figure 2 shows this equation, where $N_{scenarios}$, N_{items}, and N_{raters} are the numbers of scenarios, medical record items, and raters. h_{ij} is +1 if the search strategy would include the j^{th} item of the i^{th} scenario as a hit, and -1 if it would not be included. r_{ijk} depends on the k^{th} expert's relevance rating of the j^{th} item of the i^{th} scenario–"Certainly relevant" is scored as +1, "Probably relevant" is scored as +½, "Unlikely relevant" is scored as -½, and "Irrelevant" is scored as -1. S_{total} is normalized to range from -1 (indicating perfect disagreement between the search results and the expert relevancy ratings) through 0 (indicating no correspondence between the search results and the expert relevancy ratings), and +1 (indicating perfect agreement between the search results and the expert relevancy ratings).

A metric for the degree of concordance only for relevant included items, $S_{included}$, can also be calculated. This includes only items where $\sum r_{ijk}>0$ (that is, items rated as overall relevant). This metric indicates the extent to which search results include relevant items (ie, the "sensitivity" of the search algorithm). The opposite metric, $S_{excluded}$, which includes only items where $\sum r_{ijk}<0$, indicates the degree to which items rated as irrelevant are excluded from the search algorithm results. Both of these metrics also range from -1 to +1.

These metrics can be represented according to the open source schema, available at the referenced link [14].

Figure 2. Equation to calculate a performance score for a search strategy based off of the expert relevancy ratings.

$$S_{\text{total}} = \frac{\sum_{i=1}^{N_{\text{scenarios}}} \sum_{j=1}^{N_{\text{items}}} h_{ij} \sum_{k=1}^{N_{\text{raters}}} r_{ijk}}{\sum_{i=1}^{N_{\text{scenarios}}} \sum_{j=1}^{N_{\text{items}}} \left| \sum_{k=1}^{N_{\text{raters}}} r_{ijk} \right|}$$

Discussion

Implications

Federal subsidies in the Health Information Technology for Economic and Clinical Health Act have essentially ensured that EMR will become commonplace in US health care facilities [15]. While capturing and presenting medical information in an electronic format is an important first step, the next steps in making this information useful will include the capability to perform a simple keyword search, followed by the development of more complex, context-specific searches [16,17]. Designing context-specific searches that are both accurate and complete is particularly challenging given the large amount of unstructured free text data within the medical record [4,16]. For instance, free text patient data may be characterized by abbreviations, synonyms, acronyms, negative forms of key terms, or misspellings, all of which must be incorporated into optimized search strategies [4]. Because of these recognized pitfalls in EMR search algorithm design, it is essential to be able to quantitatively judge and refine search algorithms in a cyclic iterative development pattern.

The process described herein would allow for the use of an interactive search strategy design tool. After loading the sample medical record data and relevance ratings, the designer could modify a search strategy's metadata conditions and regular expressions and assess the overall performance changes. The tool could also be engineered to allow the designer to drill into the result set to find exemplars of the items that result in a mismatch of relevance ratings and search results. When an optimized search strategy is found, it is essentially prevalidated.

Process Advantages

One advantage to the process outlined above is that by basing the sample data on real patient medical records and physicians' specific impressions of which items are useful in a particular context, a very specific, detailed model of relevance is created which simple search heuristics are unlikely to capture well. As search strategy developers add complexity to their tools, they will be able to tell whether modifications are actually resulting in better matching.

The datasets and relevance tools can be shared, and even made semipublicly available. The universally unique identifiers attached to the scenario families, scenarios, medical record items, and raters minimize the chances of duplicated data. Individual sites can add their own patient data to already specified clinical scenarios and recheck performance given their site-specific sample data. Adding new raters and incorporating their responses can reduce the effect of individual raters' idiosyncrasies. The library of clinical scenarios can be expanded over time and shared.

The initial conception of the tool was to aid radiologists who desire relevant medical record information at the time of interpretation. However, many medical practitioners would benefit from having relevant items in the medical record highlighted for them, especially if the tool's accuracy for including relevance and excluding irrelevance is high. Additionally, these context-specific search strategies represent potentially powerful research tools, specifically related to outcomes tracking [6].

Process Limitations

There are many limitations to the search strategy validation process as described. First, the process of collecting the expert relevancy ratings is only semiautomated and therefore time intensive. Collection of the data requires clinical personnel, many of whom are already stretched thin and working in an atmosphere of shrinking margins, to take time away from clinical duties to perform the relevancy rating. The long-term viability of this semiautomated process requires further study and continuous process improvements to reduce the impact on experts. Second, the process relies completely on relevancy ratings communicated using a nondichotomous, ordinal scale of values. As a result, the method of data collection and subsequent validation framework fails to capture potentially valuable qualitative feedback from expert raters. Potential future work can be aimed to provide further nuance to the validation framework by incorporating qualitative feedback, such as free text entries from expert raters. Last, since this work only proposes and lays out this process, future work will be needed to validate the method of calculating the performance score and to determine whether search strategies validated by the process are actually deemed as useful by clinical providers in their daily practice. We expect that this mode of calculated search strategy performance will be only one component of evaluating and improving search strategies. Other important metrics of performance as well as the subjective experience of the returned results should also be considered to evaluate automated search strategies deployed for clinical use.

In this paper, we have outlined a process for developing and validating context-specific search strategies based on context-specific expert relevancy ratings. Since both the method for collecting the expert relevancy ratings and the framework for validating search strategies are provided as open-source tools with open formats for data interchange, any research group or commercial entity can develop software to bring data into this process and perform the proposed steps. We anticipate that the formats and process will be further refined over time as it is adapted to new tasks and clinical applications.

Conflicts of Interest

None declared.

References

1. Ahmed A, Chandra S, Herasevich V, Gajic O, Pickering BW. The effect of two different electronic health record user interfaces on intensive care provider task load, errors of cognition, and performance. Crit Care Med 2011 Jul;39(7):1626-1634. [doi: 10.1097/CCM.0b013e31821858a0] [Medline: 21478739]
2. Arenson RL. Recommendations for additional imaging in radiology reports: Radiologists' self-referral or good clinical practice? Radiology 2009 Nov;253(2):291-292. [doi: 10.1148/radiol.2532091216] [Medline: 19864523]
3. Campbell EJ, Krishnaraj A, Harris M, Saini S, Richter JM. Automated before-procedure electronic health record screening to assess appropriateness for GI endoscopy and sedation. Gastrointest Endosc 2012 Oct;76(4):786-792. [doi: 10.1016/j.gie.2012.06.003] [Medline: 22901989]
4. Alsara A, Warner DO, Li G, Herasevich V, Gajic O, Kor DJ. Derivation and validation of automated electronic search strategies to identify pertinent risk factors for postoperative acute lung injury. Mayo Clin Proc 2011 May;86(5):382-388 [FREE Full text] [doi: 10.4065/mcp.2010.0802] [Medline: 21531881]
5. Singh B, Singh A, Ahmed A, Wilson GA, Pickering BW, Herasevich V, et al. Derivation and validation of automated electronic search strategies to extract Charlson comorbidities from electronic medical records. Mayo Clin Proc 2012 Sep;87(9):817-824 [FREE Full text] [doi: 10.1016/j.mayocp.2012.04.015] [Medline: 22958988]
6. Seyfried L, Hanauer DA, Nease D, Albeiruti R, Kavanagh J, Kales HC. Enhanced identification of eligibility for depression research using an electronic medical record search engine. Int J Med Inform 2009 Dec;78(12):e13-e18 [FREE Full text] [doi: 10.1016/j.ijmedinf.2009.05.002] [Medline: 19560962]
7. Harvey HB, Krishnaraj A, Alkasab TK. Design of a software system to collect expert relevance ratings of medical record items for specific clinical tasks. JMIR Med Inform (forthcoming) 2014. [doi: 10.2196/medinform.3204]
8. RELAX NG home page. URL: http://relaxng.org/ [accessed 2013-12-21] [WebCite Cache ID 6M2Fylhwk]
9. GitHub - scenario. URL: https://raw.github.com/talkasab/assayer_process/master/scenario.rnc [accessed 2014-02-15] [WebCite Cache ID 6NPXLQWiv]
10. GitHub – scenario family. URL: https://raw.github.com/talkasab/assayer_process/master/scenario_family.rnc [accessed 2014-02-15] [WebCite Cache ID 6NPXgeP2E]
11. GitHub – scenario family ratings. URL: https://raw.github.com/talkasab/assayer_process/master/scenario_family_ratings.rnc [accessed 2014-02-15] [WebCite Cache ID 6NPXmAgGU]
12. GitHub – search strategy scenario family results. URL: https://raw.github.com/talkasab/assayer_process/master/search_strategy_scenario_family_results.rnc [accessed 2014-02-15] [WebCite Cache ID 6NPXrUheu]
13. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960 Apr 01;20(1):37-46. [doi: 10.1177/001316446002000104]
14. GitHub – search strategy score. URL: https://raw.github.com/talkasab/assayer_process/master/search_strategy_score.rnc [accessed 2014-02-15] [WebCite Cache ID 6NPXvMOOx]
15. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010 Aug 5;363(6):501-504. [doi: 10.1056/NEJMp1006114] [Medline: 20647183]
16. Hsu W, Taira RK, El-Saden S, Kangarloo H, Bui AA. Context-based electronic health record: Toward patient specific healthcare. IEEE Trans Inf Technol Biomed 2012 Mar;16(2):228-234. [doi: 10.1109/TITB.2012.2186149] [Medline: 22395637]
17. Bashyam V, Hsu W, Watt E, Bui AA, Kangarloo H, Taira RK. Problem-centric organization and visualization of patient imaging and clinical data. Radiographics 2009;29(2):331-343 [FREE Full text] [doi: 10.1148/rg.292085098] [Medline: 19168763]

Abbreviations

EMR: electronic medical record
XML: eXtensible Markup Language

Edited by G Eysenbach; submitted 21.12.13; peer-reviewed by M Wong, I Mircheva; comments to author 12.01.14; revised version received 02.02.14; accepted 03.02.14; published 11.03.14.

Please cite as:
Harvey H, Krishnaraj A, Alkasab TK
Use of Expert Relevancy Ratings to Validate Task-Specific Search Strategies for Electronic Medical Records
JMIR Med Inform 2014;2(1):e4
URL: http://medinform.jmir.org/2014/1/e4/
doi:10.2196/medinform.3205
PMID:25601018

Original Paper

An Intelligent Content Discovery Technique for Health Portal Content Management

Daswin De Silva[1*], PhD; Frada Burstein[1*], PhD

Centre for Organisational and Social Informatics, Faculty of IT, Monash University, Melbourne, Australia

[*]all authors contributed equally

Corresponding Author:
Daswin De Silva, PhD
Centre for Organisational and Social Informatics
Faculty of IT
Monash University
900 Dandenong Road Caulfield East
Melbourne, 3145
Australia
Phone: 61 3 99032011
Fax: 61 3 9903 1077
Email: daswin.desilva@monash.edu

Abstract

Background: Continuous content management of health information portals is a feature vital for its sustainability and widespread acceptance. Knowledge and experience of a domain expert is essential for content management in the health domain. The rate of generation of online health resources is exponential and thereby manual examination for relevance to a specific topic and audience is a formidable challenge for domain experts. Intelligent content discovery for effective content management is a less researched topic. An existing expert-endorsed content repository can provide the necessary leverage to automatically identify relevant resources and evaluate qualitative metrics.

Objective: This paper reports on the design research towards an intelligent technique for automated content discovery and ranking for health information portals. The proposed technique aims to improve efficiency of the current mostly manual process of portal content management by utilising an existing expert-endorsed content repository as a supporting base and a benchmark to evaluate the suitability of new content

Methods: A model for content management was established based on a field study of potential users. The proposed technique is integral to this content management model and executes in several phases (ie, query construction, content search, text analytics and fuzzy multi-criteria ranking). The construction of multi-dimensional search queries with input from Wordnet, the use of multi-word and single-word terms as representative semantics for text analytics and the use of fuzzy multi-criteria ranking for subjective evaluation of quality metrics are original contributions reported in this paper.

Results: The feasibility of the proposed technique was examined with experiments conducted on an actual health information portal, the BCKOnline portal. Both intermediary and final results generated by the technique are presented in the paper and these help to establish benefits of the technique and its contribution towards effective content management.

Conclusions: The prevalence of large numbers of online health resources is a key obstacle for domain experts involved in content management of health information portals and websites. The proposed technique has proven successful at search and identification of resources and the measurement of their relevance. It can be used to support the domain expert in content management and thereby ensure the health portal is up-to-date and current.

(JMIR Med Inform 2014;2(1):e7) doi:10.2196/medinform.2671

KEYWORDS

health information retrieval, personalised content management, health information portal, fuzzy multi-criteria ranking, automated content discovery, data analytics, text mining

Introduction

Background

The Internet has become a key medium for audiences seeking health information resources [1]; an important contributor is health information portals. Content management (CM) in health information portals covers a broad spectrum of functions that surround the creation, discovery, distribution, consumption, and maintenance of content. A mixture of cyclic and acyclic execution of these functions is evident in both research and industrial applications. Large organizations usually follow the full cycle from content creation to maintenance, whereas specific applications focus on the advancement of a limited number of functions. Each function has its own challenges with added complexity introduced by the context of the application.

CM is a widely published topic with research conducted in knowledge management [2], Internet research [3], and information retrieval [4]. The focus of research in CM is largely influenced by its context. This context varies from enterprise level management to management of basic website content. At the enterprise level, recent advances include the ECM3 model [5], which aims to address the CM challenges by introducing stages of maturity for all enterprise documents and unstructured content. The Web content maturity model proposed by Forrester research [6] attempts to address the challenges facing an organization's Web content. It consists of 4 phases: basic, tactical, enterprise, and engagement. The focus gradually broadens through these 4 phases, starting with the basic focus of making enterprise content available online and in the final phase expanding it to providing an online channel to achieve organizational goals. The *Content Management Bible* [7] defines CM as composed of 3 phases: the first is creation or collection of content; the second is managing storage and retrieval, versioning over time, and multiple languages etc; and the third involves publication and delivery of the content.

Content discovery plays an important part in CM as a quality intensive function that also determines the level of acceptance by a target audience. For instance, low quality and irrelevant content that fails to gain attention would limit the usefulness of the entire CM process. The significance of content discovery is also evident through its contribution to a broad spectrum of technologies, including portals (enterprise, information, and community), wikis, e-commerce, and social media.

Domain expertise is integral to content discovery. The domain expert needs to be proficient in both the subject area as well as the process of acquiring content relevant to a well-defined audience. A domain expert would maintain a high degree of emphasis on the quality of content as well as the level of personalization. Quality is generally identified in terms of 4 factors: relevance, usefulness, reliability, and timeliness [8]. Personalization addresses the diverse interests, needs, and expectations of a target audience composed of several subgroups [9].

Domain experts involved in content discovery for health information portals are confronted with an exponential growth in online content. Although access to most content is simplified by the availability of search engines, the discovery of relevant, high quality content that is personalized to suit the information needs of a target audience remains a challenge. In this paper, we propose an intelligent content discovery technique to address the challenge. This paper follows the design science research process to solve this important real world problem by designing a solution (information technology artefact) in a form of an innovative automated content discovery and ranking approach for health information portals [10].

The groundwork of the technique was reported in a previous publication [11]. The technique is based on the appropriation of an existing expert-endorsed content base as a benchmark to evaluate new content with similar features and offer the new content for inclusion to the portal repository. This semi-automated technique augments the manual process of content discovery, thus addressing inefficiencies, saving human effort, and potentially reducing human error with the increasing availability of online health information.

As stated, content discovery is relevant to a wide spectrum of technologies and application areas. This paper explores content discovery in the context of smart health information portals (SHIPs).

Smart Health Information Portals

An information portal, in general, is a gateway to a diverse collection of information on a specific domain of interest. It attempts to aggregate information from multiple sources and present it in a useful form to targeted groups of users [12]. Advances in information systems coupled with the wide availability of diverse interfaces to the Internet have led to the adoption of smart technology for the development of portals. Within this context, it is pertinent to formally define a SHIP as the provision of smart technology and techniques to enhance the core capabilities of CM, content delivery, and collaboration for online health information provision [11]. The authors identify that it is not sufficient to define SHIP exclusively on its exhibiting computational intelligence features, for example, learning, reasoning, and memory. Sustainability of SHIP operation within organizational settings is crucial for its long-term viability. Hence, the issue of maintenance support becomes one of the deciding factors in the level of intelligence of a SHIP's operation.

Breast Cancer Knowledge Online [13] and Heart Health Online [14] are examples of SHIPs researched and developed at the Faculty of Information Technology, Monash University, to address the health and medical information requirements of individuals associated with breast cancer, and mental health associated with heart conditions, including patients, caregivers, family, and friends of those affected. The delivery of user-sensitive, relevant, timely, and accurate health information to the various user groups was the focus throughout the various phases of the projects. These SHIPs are implementing several novel research outcomes, for example, resource description quality criteria modelling [15], user-centric portal design [16], automated quality assessment [8], and decision support systems perspective on portals [17]. Reported experience from the development of these SHIPs clearly demonstrated the value of continuous engagement and a high degree of reliance of user

groups to identify, categorize, and describe the type of information required by relevant individuals. The resource intensity in terms of time and scarcity of relevant expertise was also highlighted by the researchers involved in these projects [17-19]. These studies reinforce the need for intelligent support for SHIP CM.

Automated content discovery, content summarization [20], dynamic ranking, user annotations, and feedback [21] are some of the enhancements to CM, which could assist in SHIP CM. Content delivery is enhanced with user profiling, geographical filtering, mobile interfaces, and device-independent content delivery. Online messaging, social networking, and discussion forums are enablers for smart collaboration. Among these features, assurance of quality of information delivery is by far the most sought after by users, and the most resource intensive from the organizational setup point of view.

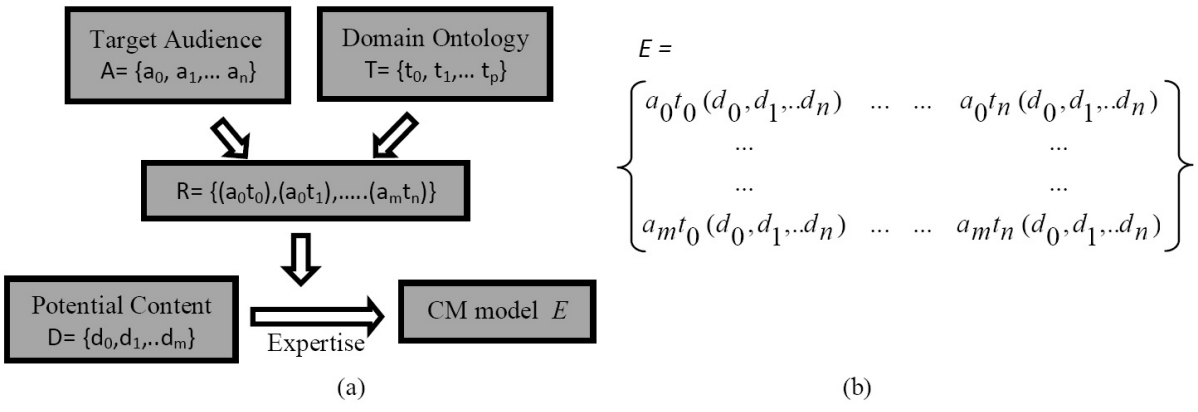
Content Management Model

The CM model represents the external entities of CM and their interactions in the formulation and management of personalized content. Informed by the experience with BCKOnline and Heart Health portal research [19], this model is a conceptualization of the fact that the audience of the SHIP users has distinct characteristics and contexts, which potentially affect their information needs. The resources for a SHIP can be aligned with a domain ontology, which classifies them against the major concepts that define such a domain. For example, official publications from medical journals are usually classified by a set of keywords, which the audience is likely to use to search and retrieve these publications. A set of such keywords or subject terms can be considered as part of domain ontology. The completeness or relevance of such an ontology can be problematic, especially when it comes to the search for relevant user-centered information [18]. It is up to the domain experts to reach consensus when deciding which terms are most suited for the ontology and content discovery. However, these issues are outside the scope of this particular paper. For this research we assume that there is a trusted and appropriate domain ontology constructed for resource classifications (eg, in BCKOnline, a combination of Medical Subject Headings [MeSH], BreastCare Victoria Glossary, BCKOnline Disease Trajectory, and BCKOnline keywords were used as encoding schemas for the subject metadata element [22]). The role of domain experts in classifying potential resources against the needs of the target audience becomes essential for identifying the best terminology suitable and understandable by the target audience.

At the generic level, the target audience, potential content, a domain ontology, and domain expertise are the external entities that are fused together to generate personalized content. This formulation is further illustrated in Figure 1a. It is useful to formally define the entities and their interactions. The target audience comprises subgroups of users with similar characteristics and thus having similar information needs. Let $A=\{a_0, a_1,\ldots a_n\}$ be the target audience comprising all subgroups. Let $D=\{d_0, d_1,\ldots d_m\}$ be the set of all content that is able to address the information needs of the target audience. A domain ontology formalizes the concept hierarchy of knowledge for a specific domain, and it can be generally represented as a set of topics, $T=\{t_0, t_1,\ldots t_p\}$. The information requirements for audience A are determined using the Cartesian product of A and T. Let R be the Cartesian product, $R=A*T$. Actual information requirements could very well be a subset of R because all terms may not be applicable to all A. Domain expertise transforms information requirements R, to actual content D, by determining subsets of D that address each R. Let this transformation be $E=\{e_0, e_1,\ldots e_x\}$, where $e_0=\{a_0t_0,(d_0,d_1,\ldots d_m)\}$ comprises information requirements and a set of matched content elements. The transformation E represents the CM model because it captures all entities and their relationships. It can also be depicted as a matrix (Figure 1b).

The CM model possesses certain properties that make it robust and flexible to changes. Over time, it is likely A, T, and D would expand or contract to reflect developments in health practices. Matrix E is time-invariant and thus can be altered easily to reflect these changes. The challenge and opportunity for developing a sustainable CM model is in designing transformation R as a semi-automated expert-driven procedure by using intelligent technologies. The following section elaborates on this technique.

Figure 1. (a) Formulation of content management entities (b) SHIP content model as a matrix.

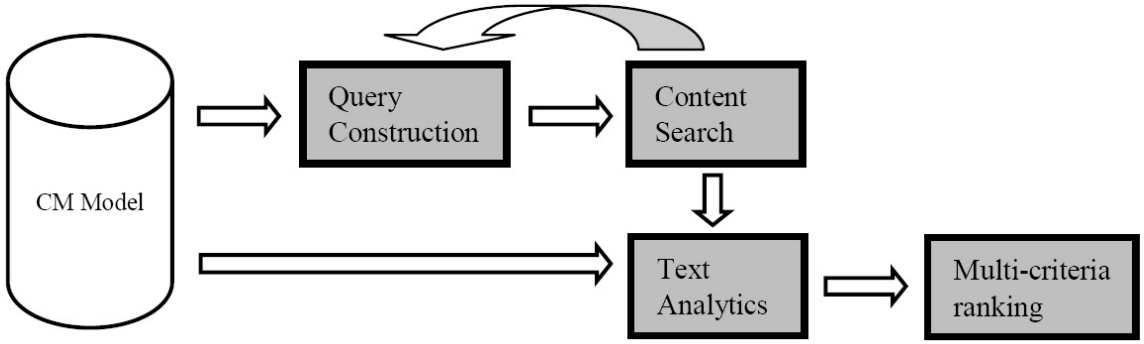


Methods

Overview

The CM model underlies the formulation of the proposed technique. It extracts semantics that are useful to construct queries that discover new content as well as semantics that are used to measure the relevance of new content from the CM model. Query construction introduces context specific information to the final query that is then distributed to search engines. The results are amalgamated and followed by the analysis of textual content of both new and existing resources. In the content selection phase, each item is ranked based on several factors of quality and presented to the domain expert for further perusal and possible inclusion in the content repository. Figure 2 illustrates the components of the technique.

Figure 2. Proposed content discovery technique.



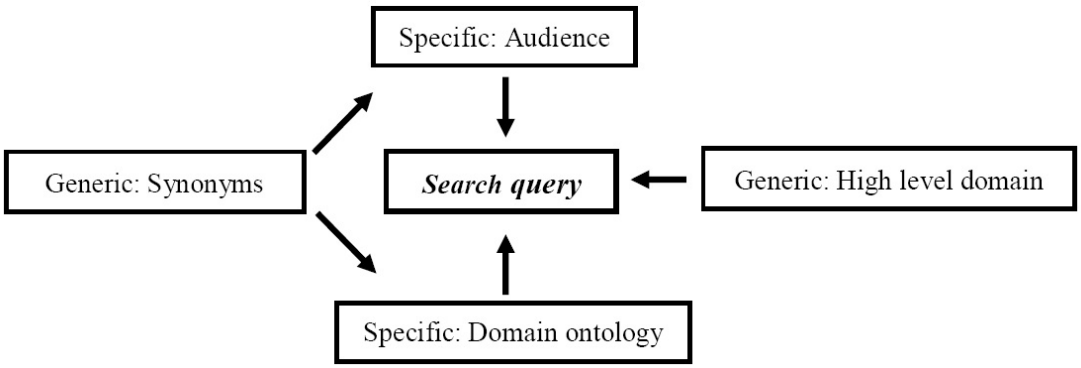
Query Construction and Content Search

Each query is based on several specific and generic dimensions. The specific dimensions are sourced from meta-data found in the first element of each term in the CM matrix (Figure 1b). The element $a_x t_y$, denotes the audience subgrouping and the term (or topic) from the domain ontology. The generic dimensions serve the purpose of introducing the context/background to a search. These can range from the high-level domain terms to synonyms indicative of the specific dimensions. Figure 3 illustrates this further.

Both specific dimensions are well defined by the domain expert and thereby translate easily into query construction. The audience dimension will contain information about the subgroups found within. Age, sex, marital status, occupational status, and level of knowledge of the domain are some examples. The domain ontology contains the key terms and phrases that define the information needs of the audience. The generic dimension of synonyms introduces further diversity to the query construction process with related terms for the two specific dimensions. The widely used lexical database, WordNet [23] is used to extract synonyms with semantic relationships. WordNet is a lexical database for the English language. It is made up of two parts: sets of synonyms called (synsets) and the semantic relations between these sets. The semantic relations are useful to identify terms that have a common ancestor and thus can be linked to each other. For instance, wellness and well-being are terms similar in meaning to health but positioned at different levels on WordNet. Query construction will generate a set of queries $Q=\{q_1,q_2,\ldots q_n\}$, representing the information needs expressed in the CM model.

Query construction and content search are recurrent phases in which queries with failed searches are reconstructed using synonyms from WordNet. In the content search phase, each query will be run on several search engines. Duplicates are removed from the search results generated and merged into one distinct set. The actual webpages are downloaded from this list and further examined for misrepresentations, such as duplicates, revisions of the same page, index pages, pages generated by other search engines, etc. The valid results are converted to plain text using Apache Tika, which is able to parse most Web document formats, including HTML, PDF, and XML. The resultant corpus of plain text documents, $D_q=d_{q1}, d_{q2}, \ldots d_{qn}; \forall n \in N, \forall q \in Q$, is input to the text analytics phase.

Figure 3. Elements in query construction.



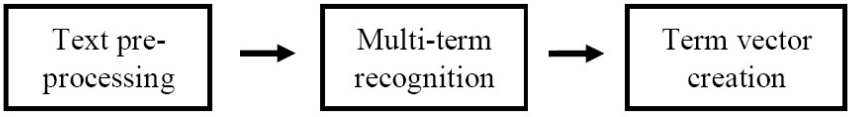
Text Analytics

Overview

Text analytics is responsible for the identification of content that is relevant to the existing expert endorsed resources. It is the core function of the technique and is made up of 3 submodules as illustrated in Figure 4.

Text analytics is an emerging area in business analytics where smart techniques are being developed and used to extract patterns, predictions, and semantic content from text corpora [24]. Every document has a number of words used only for grammar and presentation and not directly related to content description. Preprocessing removes the words that do not have a semantic use for analysis. Stop-word removal [25] and Porter's stemming algorithm [26] are run on the text corpus to generate a "bag of words" representation of each document. Further preprocessing can be conducted depending on the content of the original documents (formulae, images, and other media).

Figure 4. Text analytics sub-modules.



Multi-Term Recognition

Multi-term recognition aims to improve the semantic representation of the original document with the extraction of multi-word terms by means of the C-value/NC-value approach [27]. This method combines linguistic and statistical information with emphasis on nested multi-word terms and the general distribution of candidate terms. It has been used successfully in a variety of applications [28,29]. It generates a list of multi-word terms ranked by the NC-value. The NC-value is a weighted summation of context information and the C-value (Figure 5).

The 2 factors of NC-value have been assigned the weights 0.8 and 0.2, respectively, based on previous experiments [27]. The C-value is a measure of each term's distinct frequency of occurrence within the corpus. It takes into account the number of times the term appears nested within other candidate terms; this is subtracted from the total frequency in the corpus (Figure 6).

To improve the detection of multi-word terms, the C-value/NC-value approach was extended with the introduction of domain-specific information to the calculation of NC-value. The presence/absence of terms from the domain ontology was incorporated as shown in Figure 7.

The domain ontology is composed of terms recommended by the experts and thus would appropriately narrate the context of the search to each document. The new element in the equation captures the likelihood of candidate terms appearing within the domain ontology as nested or distinct terms. The weight of term t can be determined by the hierarchical organisation or its relationships within the ontology. The factors of the new NC-value have been assigned weights 0.6, 0.2, and 0.2, respectively This adjustment ensures that context factor and ontology information have equal contribution toward the final measure.

Figure 5. Calculation of NC-value.

$$NC\_value(a) = 0.8C\_value(a) + 0.2 \sum_{b \in C_a} f_a(b)w(b)$$

a is the candidate term,
C_a is the set of distinct context words of a,
$f_a(b)$ is the frequency of b as a term context word of a,
$w(b)$ is the weight of b as a term context word.

Figure 6. Calculation of C-value.

$$C\_value(a) = \begin{cases} \log_2|a| \cdot f(a), & distinct\ a\ (not\ nested) \\ \log_2|a| \cdot f(a) - \dfrac{1}{P(T_a)} \sum_{b \in T_a} f(b) \end{cases}$$

a is the candidate string,
$f(.)$ is frequency of occurrence in the corpus,
T_a is the set of extracted candidate terms that contain a,
$P(T_a)$ is the number of these candidate terms.

Figure 7. Calculation of NC-value with introduction of domain-specific information.

$$NC\_value(a) = 0.6C\_value(a) + 0.2 \sum_{b \in C_a} f_a(b)w(b) + 0.2 \sum_{t \in D} f_t(a)w(t)$$

D is the set of terms in the domain ontology,
$f_t(a)$ is the frequency of a as a term/nested term in an ontology term t,
$w(t)$ is the weight of t within the ontology.

Term Vector Creation

The third submodule, term vector creation, generates a vector space model (VSM) representation of the document corpus as well as the benchmark resource set. The VSM introduced by Salton et al [30] models documents as elements in term space. The term space is composed of all unique terms in the document collection and each document is represented by the vector of terms found in the document. Thereby the documents are comparable within the corpus and with external content. VSM has been successfully applied to several text mining/business analytics applications such as ontology-based information retrieval [31], incremental learning from text [32], and disease identification [33]. The VSM follows a term weighting scheme to improve the semantic position of a document. The 3 main factors of term weighting are term frequency factor, collection frequency factor, and length normalization factor. Term frequency factor determines the frequency within a single document, collection frequency factor determines its prevalence within the collection of documents, and the length of each document is used as a normalization factor to negate the bias of long documents.

A noted weakness of VSM is the assumption that identified terms are independent of each other. This shortcoming is offset to a certain degree with the inclusion of multi-word terms. Multi-word terms are able to capture more semantics than a single term set. The general VSM only focuses on single terms; therefore, it is necessary to create a separate VSM for multi-word terms. Thereby two VSMs ($vsm_m(d_q)$, $vsm(d_q)$) are created for each document d_q in each collection D generated by query q.

The VSMs generated for the document corpus need to be evaluated for relevance to the target audience and their information needs. Resources in the expert-endorsed content repository are the most suitable benchmark for this purpose. Independent to the VSMs from the document corpus D_q, separate VSMs need to be generated for these resources in the content repository. The same query sent into the content search phase is run on the content repository to identify the relevant documents, $R_q=r_1,r,\ldots r_n \; \forall q \in Q$. The content of the documents in this set is converged into a single representative document and this is sent through to the multi-word term recognition phase followed by the generation of VSMs for both multi-word terms and single terms, $vsm_m(R_q)$ and $vsm(R_q)$, respectively. The outcome from this submodule is, for each query, a set of VSMs that represent new documents found in the content search phase and a set of VSMs that represent existing resources that are have been determined by the domain expert to be relevant to the same query. Effectively, this produces a benchmark term vector and the VSMs for multi-term words, $vsm_m(R_q)$ and $vsm_m(d_q) \; \forall d \in Dq$, as well as for single terms $vsm(R_q)$ and $vsm_m(d_q)$, $\forall d \in D_q$. Both these are defined using related dimensions that enable comparisons as well as rankings.

The cosine coefficient similarity measure, which measures the angle between two vectors without bias for the length of the document, can be used to determine the closeness of each d_q to R_q (Figure 8). The denominator length-normalizes the vectors, ensuring the two are comparable in their original format. The same measure is calculated for the multi-term VSMs.

Figure 8. Calculation of cosine coefficient similarity measure.

$$cosine(R_q, d_q) = \frac{vsm(R_q) \cdot vsm(d_q)}{|\, vsm(R_q) \,||\, vsm(d_q)|}$$

Multi-Criteria Ranking

Thus far, the technique has generated 3 quantifiable measures: the ranking from content search, cosine similarity for multi-term words, and cosine similarity for single terms. Each measure represents an independent aspect of the content discovery process. The ranking from content search indicates the position assigned by the search engine (determined by the respective search and indexing algorithms) as well as its temporal significance. On the other hand, the cosine similarities are entirely content-based with the multi-term VSM capturing more semantics.

From a CM perspective, the quality of content is largely determined by 4 criteria; relevance, reliability, timeliness, and usefulness [8]. These can be defined briefly as relevance to the search query, usefulness to the target audience, reliability of the author/publishing body, and timeliness as the period when the article was compiled and published. As mentioned in the technique thus far, the existing content repository makes a significant contribution toward the relevance factor of new content. The content-based similarity measures are sound candidates for the measurement of relevance. Ranking from content search maintains temporal significance. This can be coupled with the actual date of publication, which can be retrieved from the host site (if available) to create a measure of timeliness. The author/publishing body of new content can be directly validated against authors/publishers of similar content found in the repository so that reliability can also be established

to some extent. Usefulness that cannot be determined without user involvement/feedback is the only measure of quality that is beyond the proposed content discovery technique. The quality criteria are shown in Table 1.

Table 1. Means of quality measurement derived from the technique.

Quality criteria	Means of measurement
Relevance 1	Multi-word term similarity measure
Relevance 2	Single-word term similarity measure
Reliability	Direct validation of author/publishers with existing content
Timeliness	Content search ranking and date of publication
Usefulness	Not measurable (requires target audience involvement)

Multi-criteria decision-making (MCDM) involves the identification of an alternative from a finite set based on the evaluation of values from a set of criteria that characterize the alternative [34]. Ranking of new content is a variation of MCDM where more than one alternative is selected from a set of resources based on the assessment of four factors of quality. Several methods have been proposed to address MCDM problems: crisp methods such as multiplicative exponential weighting, simple additive weighting, analytic hierarchy process [35], discrete choice analysis [36], data envelopment analysis [37], and fuzzy MCDM analysis. Fuzzy MCDM analysis is largely based on the decision-making method in a fuzzy environment developed by Bellman and Zadeh [38]. The measures of quality will reflect varying degrees of importance for each ontology term. Given this subjective nature of the qualitative factors, it is pertinent to use fuzzy MCDM analysis for selection of new content.

An MCDM problem consists of 5 elements: alternatives, criteria, outcomes, preferences, and information [39]. In the context of content ranking, the alternatives are the new content discovered, the criteria are the measures of quality, preferences are the expectations for each criterion, and the quantified measures contain the information used to evaluate these parameters. The preferences, expectations for each criterion, are subjective because they vary between terms in the domain ontology. For instance, the measure of timeliness may not be as important as relevance for certain areas of the domain that are highly theoretical with less change over time. In such cases, the outcomes can be misleading if timeliness is equally represented as relevance in the ranking scheme. In essence, the criteria are sensitive to the type of term that is being evaluated. Fuzzy MCDM analysis is advanced to overcome this limitation. The advantage of using a fuzzy approach is in the assignment of relative importance of criteria using fuzzy numbers instead of crisp numbers.

Fuzzy triangular numbers (FTN) are necessary to establish fuzzy weights for each criterion. Input provided by domain experts on the expectations of each criterion for each term is represented as FTNs. An FTN is defined as a fuzzy set, $F=\{(x,\mu_F(x),x\in R)$, where x takes values on the real line, $R:-\infty \ll x \ll \infty$ and $\mu_F(x)$ is a continuous mapping from R to closed interval [0,1]. A FTN denoted as $M=(l,m,u)$, where $l \ll m \ll u$, expresses the relative strengths of each pair of elements in the same hierarchy. The parameters l; m; u; represent the smallest possible value, the most promising value (modal), and the largest possible value respectively in a fuzzy event. The membership function of M is expressed as follows (Figure 9).

The first 4 criteria (Table 1) relevance 1, relevance 2, reliability, and timeliness are defined as $C=\{c_1, c_2, c_3, c_4\}$ respectively. The weight of criterion c assigned to term t by expert M_p is denoted as FTN: $w^p{}_c=(l^p{}_c,m^p{}_c,u^p{}_c)$, where $c \in \{c_1, c_2, c_3, c_4\}$ and $p=1,\ldots P$. The geometric mean is used to determine the aggregate weight when multiple experts provide input on expectations. The fuzzy score for criterion c of candidate resource r in terms of FTNs given by expert M_p is denoted as $s^p{}_{cr}=(LE^p{}_{cr}, ME^p{}_{cr}, UE^p{}_{cr})$ where $r=1,\ldots m$, and $P=1,\ldots PP$. An FTN for the weights of each criterion can thus be defined as $(m^p{}_c-\rho, m^p{}_c, m^p{}_c+\rho)$, where $m^p{}_c$ is the FTN mean and ρ is its spread, which is determined by domain experts and reflects the characteristics of criterion c. With R alternatives and C criteria, the weighted sum is derived to measure performance and shown in Figure 10.

Ranking takes place when $n_i > n_j$ if and only if $e_{ij}=1$ and $e_{ji}<Q$, where Q is a fixed position fraction of a number less than 1 (preferably 0.9). The use of a fuzzy MCDM approach has thus converted measures representing different qualitative factors into a single ranked metric based on weights indicative of the term from the domain ontology that is being explored by the technique. The ranked resources can now be easily perused by a domain expert.

Figure 9. Calculation of membership function of M.

$$\mu_M(x) = \begin{cases} \dfrac{x-l}{m-l} & x \in [l,m], \\ \dfrac{u-x}{u-m} & x \in [m,u], \\ 0 & otherwise. \end{cases}$$

Figure 10. Calculation of derivation of weighted sum to measure performance (top equation). Once the weighted sum has been calculated, resources can be ranked (bottom equation).

$$P = \sum_{c=1}^{K} a_{cr} \acute{W}_c$$

where a_{cr} represents the performance value of the r-th alternative in terms of the c-th criterion, with $\acute{W}_c$ representing the aggregate weights of criterion c as an FTN, $\acute{W}_c = (\acute{w}_{cl}, \acute{w}_{cm}, \acute{w}_{cu})$.

$$e_{ij} = \max\{min(\mu_i(x), \mu_i(y))\} \quad x \geq y;\ \forall\, i,j = 1,2,\ldots,m,$$

where $\mu_i(x)$ denotes the membership function for fuzzy number $\acute{n}_i$.

Results

As outlined earlier, SHIP was selected as the application test bed for the delineated technique. The technique was implemented using Java programming language for use in the experiments. Quality is essential for health information delivery and therefore maintenance and regular update of content is crucial for long-term value of the portal. The rate of generation of new health-related content far exceeds the numbers that can be manually examined by domain experts for relevance to a specific topic and audience. In this context, the benefits gained from the said technique are substantial. One of the portals noted earlier, BCKOnline, was used in this experiment. BCKOnline is a SHIP designed and developed at Monash University for the provision of personalized health information on breast cancer. A robust CM model was used by the domain experts to manage and revise the content in BCKOnline.

The evaluation sample consisted of all content in the BCKOnline portal, a domain ontology comprising 795 terms and a content repository with 900 documents. Terms were selected from the ontology for demonstration of each phase. Each document was linked to one or more ontology terms by a domain expert. Figure 11 presents the top 30 domain ontology terms in the content repository. The graph exhibits a long tail, where a larger number of the resources are categorized in smaller groups. This signifies the breadth of health information for breast cancer accessible via the portal and further justifies the need for an automated content discovery process. The highest numbers of resources are on the primary subtopics of early, advanced, and recurrent breast cancer.

"Palliative care," which has a count of 52 resources, was selected to demonstrate the query construction component. Construction of the query involves generic and specific dimensions (Figure 3). The actual term is the specific ontology dimension and the term "breast cancer" represents the high-level domain and its inclusion introduces a background to the query. The next level of construction expands the query to include personalization and diversification efforts. The audience dimension is represented using several attributes specific to the high-level domain of breast cancer. These are level of knowledge, age groups, stage of illness, and user role. WordNet is explored in search of the generic dimension of synonyms. The two terms, "palliative" and "care" are searched separately. The WordNet senses metric is used to select synonyms with a higher relevance to the input term. The association of dimensions for the said term is tabulated in Table 2. Starting with the base query "*palliative care breast cancer*," the search is gradually expanded to include the audience attributes and the synonyms. Thereby, the recurrent phases of query construction and content search contribute toward good coverage of available online content.

Figure 11. Top 30 domain ontology terms in BCKOnline.

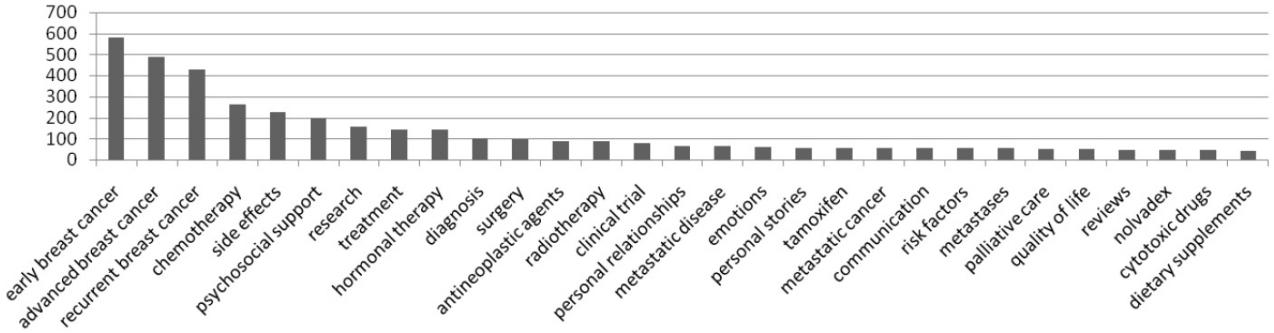


Table 2. Dimensions of query construction for term "palliative care."

Dimension	Values
Specific: audience	(basic, scientific, experiences), (young, middle-aged, old), (early, recurrent, advanced stages), (friend, partner, child)
Specific: domain ontology	palliative care
Generic: high level domain	breast cancer, breast carcinoma
Generic: synonyms: palliative	Directly related: alleviative, preventative, lenitive Inherited from: curative, remedial, therapeutic
Generic: synonyms: care	Directly related: aid, attention, tending Inherited from: work, action, procedure

After the search results have been processed into a corpus of plain text documents, $D_q=d_{q1}, d_{q2}, \ldots d_{qn}$, multi-term recognition takes place. As mentioned earlier, this module identifies multi-word terms that are ignored by the VSM. The expectation of text analytics phase is to capture semantics representative of the documents; the inclusion of multi-word and single-word terms reinforces the VSM outcomes. As an illustrative example, some comparable multi-word terms and single-word terms recognized from a high ranked resource are presented in Table 3.

Table 3. Comparison of multi-word and single-word terms from an online resource on "palliative care"[40].

Multi-word terms	Single-word terms
palliative care, palliative care team, palliative care specialist, palliative medicine, anticipate future issue, spiritual care, outpatient setting, treatment option, family member	palliative, care, specialist, treatment, disease, female, support, family, body, medicine

In the term vector creation stage, VSMs for multi-term words, $vsm_m(R_q)$ and $vsm_m(d_q)$ $\forall d \in Dq$, as well as for single terms $vsm(R_q)$ and $vsm_m(d_q)$, $\forall d \in Dq$ are generated. Vector R_q represents the benchmark vector derived from existing resources in the content repository. The cosine similarity was used to measure likeness between the VSMs with the threshold set at 0.75. Two terms were selected to demonstrate the measures of similarities. These are "palliative care" and "reviews." The contrasting nature of the terms, the first being specific and the second more general, appeals to the usual content discovery requirements of information portal and related Internet technologies. The number of new resources above the threshold for the first term was 45 and 70 for the second term. The second term, "reviews" has a larger number of resources because it covers a broad content area. The cosine similarities in the range of 0.75-1 in bins of 0.05 are depicted in the histograms in Figure 12 for the multi-word and single-word VSMs of the two terms. The primary observation here is the high similarities of most resources in the multi-word VSM, with 60 resources (23 for palliative care and 37 for reviews) in the range of 0.9-1.0 in comparison to single-word terms that have only 25 in the same range. This proximity to the benchmark is indicative of the contextual information captured by multi-word terms.

Multi-criteria ranking aims to satisfy 3 criteria: relevance, reliability, and timeliness. The multi-word and single word similarity measures make up 2 relevance measures. The ranking from the content search is coupled with the upload date and time of each resource to calculate a timeliness measure. Reliability is determined by comparing the author/publisher names of new resources with those already in the repository. Unknown authors are ranked very low so that domain experts can intervene at the actual content selection phase to determine reliability based on their knowledge. As already presented, the varying level of importance of criteria for each term prompted the use of fuzzy weights per criterion per term. Inputs

accumulated from domain experts are accumulated and aggregated to generate these FTNs. The following FTNs (Table 4) were used for the 2 terms "palliative care" and "reviews" to demonstrate the multi-criteria ranking process. Both terms have high weights for the 2 relevance measures and reliability in contrast to timeliness. Timeliness is not crucial for the term "reviews" due to the obvious nature of a medical review. The reliability measure for "review" is weighted above that for "palliative care." The weighted sum value, $a_{cr}W_c$ for three resources for term "reviews" is presented in Table 5. The 4 measures for each resource were normalized to 1-10 and are shown in the first column of Table 5.

Figure 12. Histograms of similarities of new resources to benchmark VSM (a) single-word (b) multi-word terms.

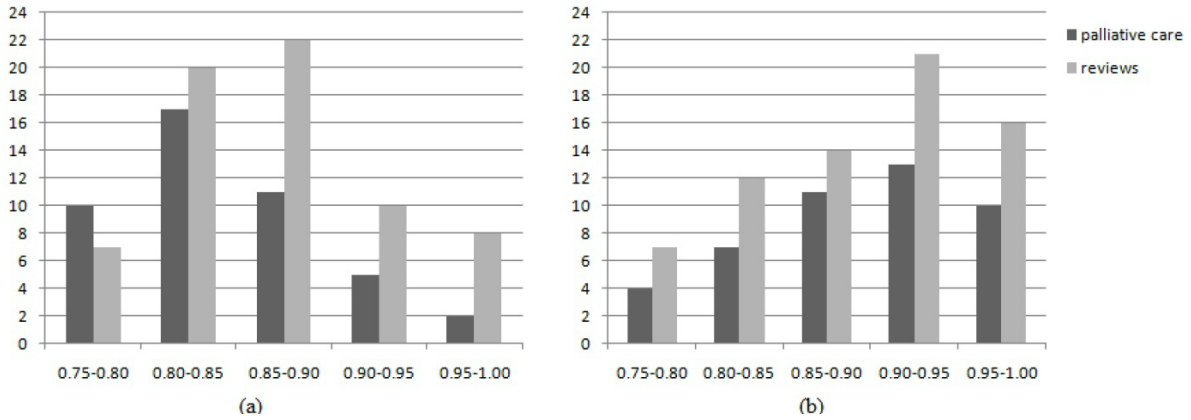


Table 4. FTNs used for ranking criteria.

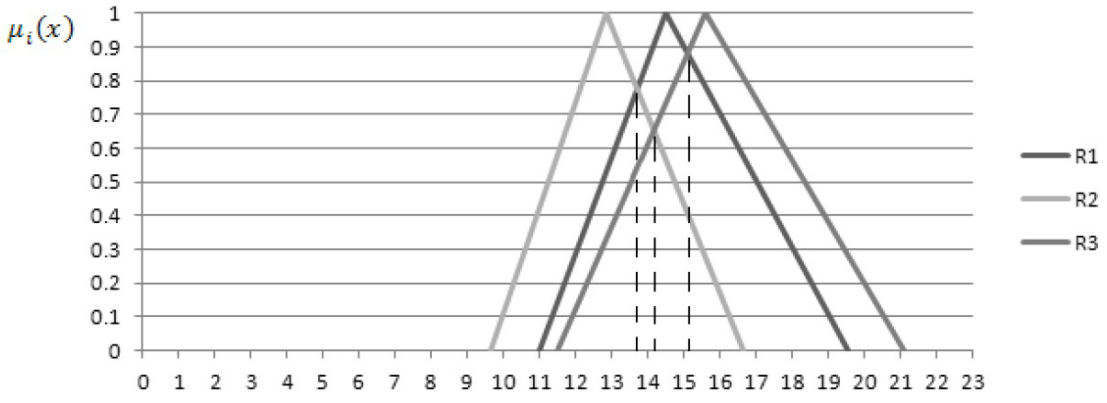
Term	Relevance 1	Relevance 2	Timeliness	Reliability
Palliative care	(0.50, 0.70, 0.90)	(0.30, 0.50, 0.70)	(0.40, 0.60, 0.70)	(0.40, 0.50, 0.60)
Reviews	(0.60, 0.70, 0.90)	(0.60, 0.70, 0.90)	(0.10, 0.30, 0.40)	(0.40, 0.60, 0.90)

Table 5. Weighted measures for three resources for term "review."

Resource name and measures	Relevance 1 (0.60, 0.70, 0.90)	Relevance 2 (0.60, 0.70, 0.90)	Timeliness (0.10, 0.30, 0.40)	Reliability (0.40, 0.60, 0.90)
R1 (7.50, 5.50, 4.10, 6.90)	(4.50, 5.25, 6.75)	(3.30, 3.85, 4.95)	(0.41, 1.23, 1.64)	(2.76, 4.14, 6.21)
R2 (5.40, 9.20, 8.70, 0)	(3.24, 3.78, 4.86)	(5.52, 6.44, 8.28)	(0.81, 2.43, 3.24)	(0,0,0,0)
R3 (8.50, 4.70, 6.80, 7.20)	(5.1, 5.95, 7.65)	(2.82, 3.29, 4.23)	(0.68, 2.04, 2.72)	(2.88, 4.32, 6.48)

The weighted summation of the resources are R1 (10.97, 14.47, 19.55), R2 (9.63, 12.83, 16.62), and R3 (11.48, 15.6, 21.08). Figure 13 displays the membership functions for each. Following Figure 10, the comparison scores are e_{31}, e_{32}, e_{12}=1, e_{13}=0.88, e_{21}=0.76 and e_{23}=0.64. Using a threshold Q of 0.9 and 0.8, respectively, the ranking of the 3 resources in descending order can be determined as R3, R1,and R2. With completion of the ranking phase, the ranked resources and the intermediary metrics are sent through to the domain expert for further scrutiny.

Figure 13. Membership functions for weighted summations of R1, R2 and R3 metrics.



Discussion

Evaluation and quality of content become crucial based on the information expectations of the target audience, especially in the case of health information [1]. The increase in relevant online health information is a challenge for domain experts to peruse and evaluate on a regular basis. This paper reported the development of an intelligent content discovery technique that is able to address this challenge with automated discovery and ranking features. The technique utilizes an existing content repository as a benchmark to validate new content discovered online. It operates in 4 modules: query construction, content search, text analytics, and multi-criteria ranking. Query construction uses an existing ontology of key terms and supplements this with audience and context information as well as synonyms extracted from WordNet. Content search retrieves a unique list of resources that are downloaded, preprocessed, and consumed by text analytics. Semantics, based on multi-word and single-word terms, are identified in text analytics and used to measure proximity to a benchmark vector derived from existing content. Acknowledging the subjective nature of qualitative factors, fuzzy weights are used in the multi-criteria ranking phase to determine a single rank encompassing relevance, reliability, and timeliness. The paper delineates the complete technique with an inclusive demonstration of its execution using an actual health information portal as a test bed. The technique can be sufficiently generalized and applied in other domains. In the next phase of the project, we will focus on validation of the technique with experiments involving domain experts as well as user studies to highlight its benefits and further establish its purpose in CM. Future research will also investigate the advantages of ripple-down rules [41] over fuzzy MCDM when generalizing the technique for application in other domains with incremental usage over time.

Acknowledgments

This research is partly supported by the University of Ballarat “Self-sustaining Regions Research and Innovation Initiative,” an Australian Government Collaborative Research Network.

Conflicts of Interest

None declared.

References

1. Eysenbach G, Powell J, Kuss O, Sa E. Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web. JAMA 2002 May 22;287(20):2691. [doi: 10.1001/jama.287.20.2691]
2. Bonifacio M, Franz T, Staab S. A four-layer model for information technology support of knowledge management. In: Becerra-Fernandez I, Leidner DE, editors. Knowledge management: an evolutionary view. Armonk, N.Y: M.E. Sharpe, Inc; 2008:104-124.
3. Tatnall A. Portals, portals everywhere ...web portals. In: Tatnall A, editor. Web Portals: The New Gateways to Internet Information and Services. Pennsylvania: Idea Group Publishing; 2005.
4. Bates M. Information Retrieval Systems. In: Understanding Information Retrieval Systems: Management, Types, and Standards. Boca Raton, FL: CRC Press; 2011.
5. Guiding Consortium. ECM3 Maturity Model. 2009. URL: http://ecm3.org/ [accessed 2014-04-15] [WebCite Cache ID 6OqfKHBmx]
6. WCMM. Web content maturity model. Forrester Research. 2009. URL: http://www.forrester.com/The+Web+Content+Management+Maturity+Model/fulltext/-/E-RES46563 [accessed 2014-04-15] [WebCite Cache ID 6OqlTnB6i]
7. Boiko B. 2001. In: Content Management Bible. Indianapolis, IN: Wiley Pub; 2005.
8. Xie J, Burstein F. Using machine learning to support resource quality assessment: an adaptive attribute-based approach for health information portals. 2011 Presented at: Proceedings of the 16th International Conference on Database Systems for Advanced Applications International Workshops; 2011; Hong Kong, China p. 526-537. [doi: 10.1007/978-3-642-20244-5_50]
9. Fisher J, Burstein F, Lazarenko K, Lynch K, McKemmish S. Health information websites: is the health consumer being well served? 2007 Presented at: Proceedings of Americas' Conference on Information Systems (AMCIS); 2007; Keystone, Colorado.
10. March S, Storey VC. Design science in the information systems discipline: an introduction to the special issue on design science research. MIS Quarterly 2008:725-730.
11. De Silva D, Burstein F. Supporting personalised content management in smart health information portals. In: Australasian Conference on Information Systems. 2012 Presented at: ACIS 2012: 23rd Australasian Conference on Information Systems; 2012; Geelong, Australia p. 1-11.
12. Breast Cancer Knowledge Online. 2014. URL: http://www.bckonline.monash.edu.au/home/welcome.jsp;jsessionid=90E8246A275FA4F8F7D951D8DD6BDA2A [accessed 2014-04-15] [WebCite Cache ID 6Oqwx8vmr]
13. Heart Health Online. URL: http://www.sip.infotech.monash.edu.au/heart-portal/home/welcome.jsp;jsessionid=B4BDBF40B30EFFF4BD9EACB469EB4237 [accessed 2014-04-15] [WebCite Cache ID 6OqwynJXL]

14. Collins H. Enterprise Knowledge Portals: Next-Generation Portal Solutions for Dynamic Information Access, Better Decision Making, and Maximum Results. New York: AMACOM; 2003.
15. McKemmish S, Manaszewicz R, Burstein F, Fisher J. Consumer empowerment through metadata-based information quality reporting: The Breast Cancer Knowledge Online Portal. J. Am. Soc. Inf. Sci 2009 Sep;60(9):1792-1807. [doi: 10.1002/asi.21096]
16. Fisher J, Manaszewicz R, McKemmish S, M., Burstein F, Malhotra P, Moon J. User-centric portal design for quality health information provision: breast cancer knowledge online. 2004 Presented at: Proceedings of the 15th Australasian Conference on Information Systems; 2004; Tasmania, Australia.
17. Burstein F, Fisher J, McKemmish S, Manaszewicz R. A role for information portals as intelligent decision support systems: breast cancer knowledge online experience. In: Intelligent Decision-making Support Systems. London: Springer; 2006:359-383.
18. Burstein F, Fisher J, McKemmish S, Manaszewicz R, Malhotra P. User centered quality health information provision: benefits and challenges. : Ieee; 2005 Presented at: Proceedings of the 38th Annual Hawaii International Conference on System Sciences; 2005; Hawaii, USA p. 1530-1605. [doi: 10.1109/HICSS.2005.651]
19. Pier C, Shandley KA, Fisher JL, Burstein F, Nelson MR, Piterman L. Identifying the health and mental health information needs of people with coronary heart disease, with and without depression. Med J Aust 2008 Jun 16;188(12 Suppl):S142-S144. [Medline: 18558916]
20. Erkan G, Radev D. LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. Journal of Artificial Intelligence Research 2004;22(1):457-479.
21. Ciccarese P, Ocana M, Garcia Castro L, Das S, Clark T. An open annotation ontology for science on web 3.0. J Biomed Semantics 2011;2(Suppl 2):S4. [doi: 10.1186/2041-1480-2-S2-S4]
22. An intelligent, user-sensitive portal to breast cancer knowledge online. Australia: Centre for Organisational and Social Informatics, Monash University URL: http://infotech.monash.edu/research/about/centres/cosi/projects/bcko/about.html [accessed 2014-04-16] [WebCite Cache ID 6Os47q1Sx]
23. Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ. Introduction to WordNet: an on-line lexical database. International Journal of Lexicography 1990;3(4):235-244. [doi: 10.1093/ijl/3.4.235]
24. Aggarwal C. In: Zhai C, editor. Mining Text Data. London: Springer; 2012.
25. Wilbur W, Sirotkin K. The automatic identification of stop words. Journal of Information Science 1992 Jan 01;18(1):45-55. [doi: 10.1177/016555159201800106]
26. Porter M. An algorithm for suffix stripping. Program: electronic library and information systems 1980;14(3):130-137. [doi: 10.1108/eb046814]
27. Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms:. the C-value/NC-value method. International Journal on Digital Libraries 2000 Aug 1;3(2):115-130. [doi: 10.1007/s007999900023]
28. van de Glind E, van Munster B, Spijker R, Scholten R, Hooft L. Search filters to identify geriatric medicine in Medline. Journal of the American Medical Informatics Association 2012;19(3):468-472. [doi: 10.1136/amiajnl-2011-000319]
29. Moschopoulos T, Iosif E, Demetropoulou L, Potamianos A, Narayanan SS. Toward the automatic extraction of policy networks using web links and documents. IEEE Trans. Knowl. Data Eng 2013 Oct;25(10):2404-2417. [doi: 10.1109/TKDE.2012.159]
30. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Commun. ACM 1975;18(11):613-620. [doi: 10.1145/361219.361220]
31. Castells P, Fernandez M, Vallet D. An adaptation of the vector-space model for ontology-based information retrieval. IEEE Trans. Knowl. Data Eng 2007;19(2):261-272. [doi: 10.1109/TKDE.2007.22]
32. De Silva D, Alahakoon D. Incremental knowledge acquisition and self learning from text. : IEEE; 2010 Presented at: The International Joint Conference on Neural Networks (IJCNN); 2010; Barcelona, Spain p. 1-8. [doi: 10.1109/IJCNN.2010.5596612]
33. Sarkar I. A vector space model approach to identify genetically related diseases. Journal of the American Medical Informatics Association 2012;19(2):249-254 [FREE Full text] [doi: 10.1136/amiajnl-2011-000480] [Medline: 22227640]
34. Brans J, Vincke P. Note—a preference ranking organisation method. Management Science 1985 Jun;31(6):647-656. [doi: 10.1287/mnsc.31.6.647]
35. Saaty T, Peniwati K. Group Decision Making: Drawing Out and Reconciling Differences. Pittsburgh, PA: RWS Publications; 2007.
36. Ben-Akiva M, Lerman SR. Discrete Choice Analysis: Theory and Application to Travel Demand. Cambridge, Mass: MIT Press; 1985.
37. Banker R, Charnes A, Cooper WW. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science 1984 Sep;30(9):1078-1092. [doi: 10.1287/mnsc.30.9.1078]
38. Bellman R, Zadeh LA. Decision-making in a fuzzy environment. Management Science 1970 Dec;17(4):B-141-B-164. [doi: 10.1287/mnsc.17.4.B141]
39. Tsiporkova E, Boeva V. Multi-step ranking of alternatives in a multi-criteria and multi-expert decision making environment. Information Sciences 2006 Sep;176(18):2673-2697. [doi: 10.1016/j.ins.2005.11.010]

40. Get Palliative Care. 2012. Breast Cancer and Palliative Care URL: http://www.getpalliativecare.org/2012/09/breast-cancer-and-palliative-care/ [accessed 2014-04-17] [WebCite Cache ID 6Ou8xgneG]
41. Yearwood J, Stranieri A. The generic/actual argument model of practical reasoning. Decision Support Systems 2006 Jan;41(2):358-379. [doi: 10.1016/j.dss.2004.07.004]

Abbreviations

CM: content management
FTN: fuzzy triangular numbers
MCDM: multi-criteria decision-making
SHIPs: smart health information portals
VSM: vector space model

Edited by G Eysenbach; submitted 16.04.13; peer-reviewed by A Stranieri, L Nguyen; comments to author 15.06.13; revised version received 04.03.14; accepted 30.03.14; published 23.04.14.

Please cite as:
De Silva D, Burstein F
An Intelligent Content Discovery Technique for Health Portal Content Management
JMIR Med Inform 2014;2(1):e7
URL: http://medinform.jmir.org/2014/1/e7/
doi:10.2196/medinform.2671
PMID:25654440

Review

Factors Associated With Adoption of Health Information Technology: A Conceptual Model Based on a Systematic Review

Clemens Scott Kruse[1], MBA, MHA, MSIT, PhD; Jonathan DeShazo[2], MPH, PhD; Forest Kim[3], MBA, MHA, Master of Arts, PhD; Lawrence Fulton[4], PhD

[1]College of Health Professions, School of Health Administration, Texas State University, San Marcos, TX, United States

[2]Department of Health Administration, Virginia Commonwealth University, Richmond, VA, United States

[3]School of Business, Baylor University, Waco, TX, United States

[4]College of Business, Department of CIS and Quantitative Methods, Texas State University, San Marcos, TX, United States

Corresponding Author:
Clemens Scott Kruse, MBA, MHA, MSIT, PhD
College of Health Professions
School of Health Administration
Texas State University
HPB, 2nd Floor
601 University Drive
San Marcos, TX, 78666
United States
Phone: 1 210 355 4742
Fax: 1 512 245 8712
Email: scottkruse@txstate.edu

Abstract

Background: The Health Information Technology for Economic and Clinical Health Act (HITECH) allocated $19.2 billion to incentivize adoption of the electronic health record (EHR). Since 2009, Meaningful Use Criteria have dominated information technology (IT) strategy. Health care organizations have struggled to meet expectations and avoid penalties to reimbursements from the Center for Medicare and Medicaid Services (CMS). Organizational theories attempt to explain factors that influence organizational change, and many theories address changes in organizational strategy. However, due to the complexities of the health care industry, existing organizational theories fall short of demonstrating association with significant health care IT implementations. There is no organizational theory for health care that identifies, groups, and analyzes both internal and external factors of influence for large health care IT implementations like adoption of the EHR.

Objective: The purpose of this systematic review is to identify a full-spectrum of both internal organizational and external environmental factors associated with the adoption of health information technology (HIT), specifically the EHR. The result is a conceptual model that is commensurate with the complexity of with the health care sector.

Methods: We performed a systematic literature search in PubMed (restricted to English), EBSCO Host, and Google Scholar for both empirical studies and theory-based writing from 1993-2013 that demonstrated association between influential factors and three modes of HIT: EHR, electronic medical record (EMR), and computerized provider order entry (CPOE). We also looked at published books on organizational theories. We made notes and noted trends on adoption factors. These factors were grouped as adoption factors associated with various versions of EHR adoption.

Results: The resulting conceptual model summarizes the diversity of independent variables (IVs) and dependent variables (DVs) used in articles, editorials, books, as well as quantitative and qualitative studies (n=83). As of 2009, only 16.30% (815/4999) of nonfederal, acute-care hospitals had adopted a fully interoperable EHR. From the 83 articles reviewed in this study, 16/83 (19%) identified internal organizational factors and 9/83 (11%) identified external environmental factors associated with adoption of the EHR, EMR, or CPOE. The conceptual model for EHR adoption associates each variable with the work that identified it.

Conclusions: Commonalities exist in the literature for internal organizational and external environmental factors associated with the adoption of the EHR and/or CPOE. The conceptual model for EHR adoption associates internal and external factors, specific to the health care industry, associated with adoption of the EHR. It becomes apparent that these factors have some level of association, but the association is not consistently calculated individually or in combination. To better understand effective

adoption strategies, empirical studies should be performed from this conceptual model to quantify the positive or negative effect of each factor.

(JMIR Med Inform 2014;2(1):e9) doi:10.2196/medinform.3106

KEYWORDS

electronic health record (EHR); electronic medical record (EMR); health information technology (HIT); medical information systems; computerized provider order entry (CPOE); adoption

Introduction

Background

The US Government passed the Health Information Technology for Economic and Clinical Health (HITECH) act [1] to incentivize adoption of the electronic health record (EHR) and to assuage the short run (SR) effects of cost to the health care organization in the adoption process. The three phases of Meaningful Use consume information technology (IT) strategies in the SR because of the HITECH act's timeline for health care organizations to qualify for monetary incentives [2,3].

Adoption of the EHR is a significant goal. International vernacular for the EHR varies; for example, electronic patient record, computerized patient records, electronic medical records (EMRs), and digital medical record. The defining difference, as defined by the Institute of Medicine, the health arm of the US National Academy of Sciences, focuses on the longitudinal and interoperable nature of the electronic patient record [4]. Without these capabilities, the patient record is greatly limited in scope. The longitudinal and interoperable nuances of the EHR are not the only significant advantages; there are eventual cost savings as well.

Studies estimate that adoption of the EHR could eventually save more than $813 billion annually, prevent 200,000 adverse drug events, and enhance the doctor-patient relationship through increased communication [5]. Unfortunately, these benefits are realized in the long run (LR), while the investment to adopt the EHR is expended in the SR. A large deficit in the SR could inhibit a health care organization's ability to compete or survive in heavily competitive environment.

The environment of health care is unique in a competitive environment. The health care organization develops an organizational strategy based on the local environment. To increase an organization's ability to compete, its strategy might also include cost reduction, and EHR adoption runs counter to this goal in the SR. The health care environment faces many sources of influence, including a reluctance to accept technology.

There has been a tremendous amount of research dedicated to the study of acceptance of technology, specifically the Technology Acceptance Model (TAM) [6]. More recent work has suggested modifications to the TAM that explain a perception of usefulness and intentions from the aspect of social influence and the cognitive instrumental process [7,8]. Several organizational theories have been developed. These focus on the sources of influence and the reason for their existence.

Organizational Theories

Organizational theories address influence, but none adequately addresses the complexity of the health care organization. Payers, providers, and patients all control resources that exert influence. The nature of the competitive environment will also exert influence on decisions. External influence from those who control resources can be explained through resource dependence theory [9,10]. Internal and external influences can be explained by the Diffusion of Innovation Theory through its introduction of compatibility, complexity, trialability, observability, and relative advantage [11-13].

According to resource dependence theory, health care organizations with the greatest level of dependence on other organizations that control the resources will feel the greatest level of environmental influence on its decisions [14]. The Resource Dependence Theory describes an external interdependence of organizations. *External Control of Organizations*, [14], which is an adaptation of Resource Dependence Theory, provides good insight for this study. The authors' premise is that the external environment creates a social context and plays an important role in how organizational decisions are made. The lack of absolute independence requires some degree of interorganizational exchange of goods or services [14]. As organizations build and negotiate relationships with each other in the exchange of resources, positions of power are established. No one organization can provide all of its own resources, so each organization becomes dependent on the other organizations that control the resources.

Similar to Resource Dependence, the Diffusion of Innovation Theory describes a social system that influences through communication channels [11-13]. Diffusion of Innovation attempts to explain how "an *innovation*, is *communicated* through *channels over time* among members of a *social system*" [13]. This theory accounts for 49-97% of variance in the rate of adoption of innovation through five factors: compatibility, complexity, trialability, observability, and relative advantage. These factors are sorted into three categories of a predictive model for EHR adoption: innovation determinants, organizational determinants, and environmental determinants [8]. The next several paragraphs exercise the five factors to this study.

The concept of compatibility [13] goes beyond answering the question, "is a product/service right for a market?" It also asks, "is the market ready for the product/service?" For instance, the Chevy Nova failed in Spanish-speaking markets because in Spanish the word "Nova" means "does not go." Promotion of conservation techniques to farmers in the United States initially failed because farmers associated conservation with lower crop

yield. Boiling water to sanitize it makes perfect sense to a market that is familiar with germ theory, but primitive tribes in Peru only heated water for sicker, weaker members; as a result, the concept failed when initially introduced and dysentery continued to flourish. In relation to this study, the concept of compatibility might ask, "is the market ready for the EHR?"

The concept of complexity [13] is appropriate to this study because innovation can be a double-edged sword. On one hand, it is new and may offer some improvement to a product or service. However, it might also be perceived as too complex; and perception can be a powerful force. If the Baby Boomer generation perceives computers to be too complex, and this perception causes computer anxiety, its users may reject its adoption and use [15]. The older physicians in a hospital have greater seniority, and are therefore, more influential in the hospital's decision to adopt the EHR. Would this same generation of providers influence the health care organization considering EHR adoption?

The concept of trialability [13] applies more to the early-adopter group than the other groups: innovators, early-majority, late-majority, and laggards. In the early phase of promotion for a new product or service, the vendor might lower the risk of adoption by offering free trials or samples to potential users. Once the user is confident of the new item's efficacy, then he/she is more likely to pay full price for its use. When a new producer of an EHR enters the marketplace, it must incentivize the use of its product because it is not known in the industry. The user accepts a risk by trying the new EHR, but the risk is overcome by the incentive. Once the new EHR gains momentum in the industry, adoption enters the early-majority phase. The new EHR has already gained momentum in the industry, and the producer does not need to incentivize its use.

The concept of observability [13] is also highly applicable to this study. Decision makers in a hospital that has not yet adopted an EHR will observe the experiences of other hospitals that have adopted it. Vendors will promote or advertise specifically to the nonadopters and help them observe how the EHR can benefit its organization. External players in the health care organization's competitive environment will provide some level of observability.

Relative advantage is a multifaceted concept for this study. In health care, the most important factor is provision of health, as well as the treatment and prevention of disease. If adoption of the EHR speaks directly to the health care organization's primary purpose, then it might provide relative advantage over competitors that have not adopted it. Another concept is that of social prestige [13]. Unless a health care organization can serve as an example to other health care organizations (observability), there may not be a sufficient level of relative advantage to be considered.

Strategy and Decision Making

Strategy can be a multifaceted concept, and organizations around the world hire strategy experts to help identify and focus on a market forces. An operational definition of strategy is borrowed from education [16] and is adapted to health care: strategy is defined as instruments by which *health care organizations* manage their organizational processes and deal with their environments in order to select a portfolio of activities and find appropriate position in the *health care industry* (italics indicate a change in wording from the authors' definition). It follows that adoption of an EHR would alter how a health care organization manages its organizational processes, so this definition of strategy is a good fit for the health care industry. However, two significant considerations in the health care environment are the level of local competiveness, and how health care organizations compete [17].

Studies have shown that decision making in the health care industry is often based on how the organization competes, whether in a single-market or multimarket environment [18]. In either environment, decision-making varies on competition, and the health care industry competes in clusters [18]. The way health care organizations compete will also affect its organizational structure. A four-cluster solution was identified as a reliable, internally valid, and stable model for health networks and a five-cluster solution for health systems [19]. Differentiation and centralization are particularly important in distinguishing unique clusters of organizations. High differentiation typically occurs with low centralization, which suggests that a broader scope of activity is more difficult to centrally coordinate. Integration is also important, but the authors find that health networks and systems typically engage in both ownership-based and contractual-based integration or they are not integrated at all.

Ash and Bates [20] studied the EHR adoption rates and the factors and forces affecting system adoption through surveys (85/650, 13.1%). Only 106 of the 650 (16.3%) of hospitals surveyed had adopted some form of EHR, 63/106 (59.4%) had implemented a full Computerized Provider Order Entry (CPOE) solution, and the other 43/106 (40.6%) implemented a partial CPOE solution. A full one-third of adopters were either Veterans Affairs or military hospitals. Additionally, 481/650 (73.8%) of those who planned to implement a full solution intended to do so within 5 years. Ash and Bates [20] also found that the size of hospital is positively-associated with component adoption; specifically CPOE adoption. The authors inferred from their results that the primary reasons to adopt the EHR is to gain the quality-of-care advantages of CPOE. This inference reinforced our inclusion of CPOE as a dependent variable.

Factors that influence health information system (HIS) adoption in US hospitals have been studied by others as well (n=1441) [21]. Results showed that HIS adoption is influenced by the hospital market, organizational, and financial factors. Larger, system-affiliated, and for-profit hospitals with more preferred provider organization contracts are more likely to adopt managerial information systems than other hospitals. Operating revenue is positively associated with HIS adoption. The study also identified hostility as an aspect of environmental uncertainty, and that organizations often turn to technological adoption to regain competitive advantage.

A knowledge-based taxonomy of critical factors for adopting an EHR was developed from a systematic literature review [22]. The researchers selected 68 of 3400 (2.00%) articles to identify six factors of adoption, listed in order of importance: user

attitude toward information systems, workflow impact, interoperability, technical support, communication among users, and expert support.

Alternative measures of EHR adoption among hospitals have been studied [23]. Authors analyzed a 2009 information technology supplement survey distributed by the American Hospital Association (AHA). The survey focused on 24 EHR functionalities in various areas: electronic clinical documentation, results viewing, CPOE, and clinical decision support. They found that 142 of 3937 (3.60%) acute-care hospitals in the United States of responding hospitals have implemented all 24 functions, 386/3937 (9.80%) of hospitals have implemented at least 20 functions, and 1437/3937 (36.50%) have implemented at least one-half of the functions. The researchers added that EHR adoption is a complex process.

Others have studied the relationship between hospital financial position and the adoption of the EHR [24]. Through a cross-sectional study of secondary data from several sources, including the AHA (2442/5752, 42.51% acute-care hospitals in the United States), researchers identified five independent and one dependent variable. Of the five independent variables (IVs), only liquidity was positively-associated with EHR. Asset turnover was negatively-associated with EHR adoption. Bed size, a control variable, was positively-associated with EHR adoption. The authors concluded that hospitals adopt EHRs as a strategic move to better align themselves with their environment.

Because commonly used elements of organizational strategy are difficult to change, several of the variables were categorized as internal organizational factors. Research has assessed variables of hospital influence in five categories: (1) capacity as measured by number of beds in groupings by intervals of 100, (2) management, or ownership, (3) organizational focus, or teaching status, (4) competitive location and alternatives, and (5) state regulatory pressures [25].

Although resources have been consumed to study factors associated with adoption of HIT, there is a gap in the literature that provides a conceptual model to guide the design of empirical models. It may seem backward to design a conceptual model after so many studies have already been conducted, but the gap remains. The aim of this study was to develop a conceptual model from a systematic literature review that associates both internal and external factors associated with adoption of the EHR. The intent of the conceptual model is to enable future empirical models.

Methods

Literature Review Process

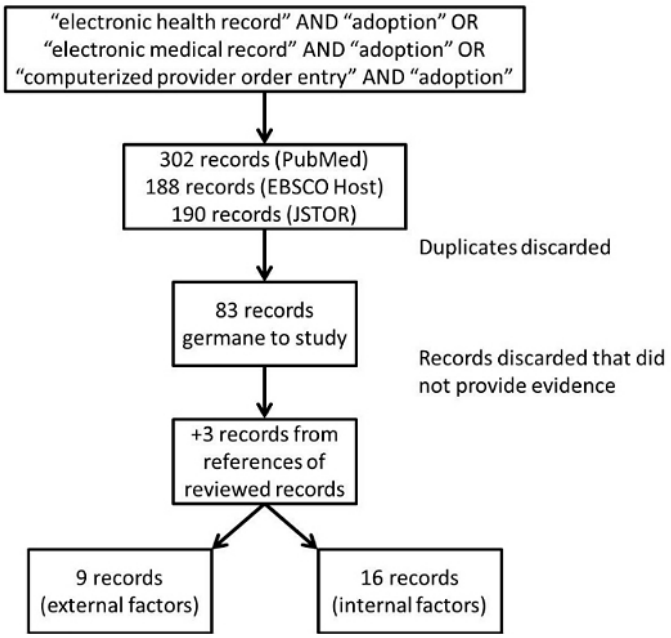
Search terms were selected based on the experience of the authors in the field of health care administration. The time frame of 1993-2013 was selected as convenience. It was assumed that 2 decades would be sufficient to capture trends.

Figure 1 illustrates the literature review process that identified 83 sources consisting of empirical studies, articles, editorials, commentaries, opinion papers, organizational theories, and text books. The intent of no limits to the type of papers was to mitigate the risk of missing something significant from a study that was not catalogued properly within a key word catalogue like the Dublin Core.

The 83 records were reviewed for content and evidence. After discarding 58 articles for lack of evidence, three additional references were added because they were key concepts upon which other studies were based. Of the remaining 25 articles, a list of factors was identified as IVs. Some factors were grouped under a similar category for the purposes of simplification of the conceptual model. The dependent variable (DV) started as adoption of the EHR, but the studies from those chosen were not as specific. From personal experience, many studies seem to discuss the EHR, but call it something else: most commonly the EMR. That is why EMR was included in the search. Because so few ERHs exist without some form of CPOE, the latter term was included in the search criteria.

Our study combines the influences highlighted by previous work and examines determinants of EHR adoption. Examining EHR adoption at the health care organization level will demonstrate validity between this study and others that have used the hospital as the unit of analysis.

Figure 1. Literature review process.



EHR Adoption and Internal Organizational Influence

Several influences in the environment exert pressure on the health care organization to adopt EHR. Influences range from incentives from the federal government to the nature of local competitive community. US federal incentives provide a heavy influence for EHR implementation, under specific conditions, and imposes penalties for a lack of EHR implementation.

The internal politics of one organization serve as one source of influence. A hospital is part of a community, which serves as an external influence. Further, if a hospital is also part of a larger multihospital system (MHS), then the politics of the broad MHS will also exert influence on local decisions.

EHR Adoption and External Environmental Influence

The patient is external to the organization, and for our study, the patient primarily serves as an external influence. Although some employees of the health care organization might also be patients, and this relationship could create a small internal influence, this study considers those few stake holders in the internal organizational factor of users. The providers are internal to the organization, and for our study, providers serve as an internal organizational influence. The payer is a significant influence [14], and the Center for Medicare and Medicaid Services (CMS) serves as a good example of this significant influence [26]. The HITECH act provides monetary incentives for EHR adoption. Those who do not implement all aspects specified in the stages of adoption are not eligible for the incentives. In this way, the CMS disincentivizes those organizations that do not adopt the EHR. If payments from the CMS were of little consequence to the health care organization's revenue, then the health care organization might decide differently about EHR adoption. A competing health care organization is an external market force in the environment. Third-party payers might compare health care organizations based on maturity of automation because mature clinical components like CPOE will result in more accurate billing. Such forces incentivize a health care organization to adopt the EHR.

Overview of the Conceptual Model

The premise for an EHR adoption conceptual model is that that environmental influences affect organizational strategy of the health care organizations that adopt the EHR [13,14,20,22,24,25]. Diffusion of Innovation theory provides three categories of a predictive model for EHR adoption: innovation determinants, organizational determinants, and environmental determinants [13]. Resource Dependence Theory provides a category of a predictive model for EHR adoption, the competitive environment. In construction of the EHR adoption conceptual model, several constructs emerged [14].

Elements of organizational strategy are not variables that can be easily changed [19]; therefore, elements typically ascribed to strategy, such as size, ownership, and fiscal stability, will be absorbed into the IVs of influence. This research proposes a model, whereby environmental factors are associated with an organization's decision to adopt the EHR.

Resource Dependence Theory explains environmental influences and the external interdependence of organizations [14]. The authors' premise is that the external environment creates a social context and plays an important role in how organizational decisions are made. The interdependence of organizations widens the field of stakeholders, and this relationship effect should be defined.

Disparate stakeholders have different interests with reference to different components of the EHR. These interests may be different in the SR interests versus the LR interests. SR interests are those that are immediate, such as current year expenditures. LR interests are further out when all inputs are variable. The SR interests of cost can often compete with the LR potential of cost savings and greater safety. Both the SR and LR interests are affected by the external environment [17].

In a highly competitive environment, SR cost implications could often win over any long-term savings. The number of patients in a market is fixed in the SR, and a highly competitive market will affect each competitor's share of that market. The SR costs of EHR implementation might be insurmountable by an organization in this market because it could not afford to lose ground without significant capital reserves or the ability to borrow cheaply [17]. However, in a less competitive market, the LR interests of potential cost savings have a better chance of influencing the decision to implement an EHR because the costs incurred in the SR are justified by the long-term benefits [17].

External stakeholders that control resources important to the health care organization can exert significant influence. For instance, a health care organization that receives a significant amount of revenue from the CMS will be influenced more by incentives provided by the CMS than an organization that receives a significant cash flow from private third parties. The relative influence of various external stakeholders may be captured by an analysis of the structure of the market in which a health care organization operates.

Stakeholders have varying interests with regard to the capabilities and effects of EHR components depending upon their relationship with the health care organization. Private payers have both SR and LR interests in the EHR. In the SR, their focus is on minimizing expenditures. Because the health care organization would pass on the implementation costs through higher contract costs, payers would not be equal in the SR. In addition, the disruption of EHR implementation could potentially affect care processes and therefore increase claims. Payers would be interested in the LR benefits of the EHR: potential cost savings, better disease management, and increased safety. However, the SR interests of the private payers might overshadow the LR benefits of the EHR. Public payers enable care of the indigent and elderly. As part of the United States Department of Health and Human Services (HHS), the CMS is highly interested in disease management, public health, safety, and research, and it may value these LR capabilities of the EHR more than the SR costs. The CMS, as part of HHS, would also favor the EHR because it supports the Presidential directive to promote the establishment of the Nationwide Health Information Network that links electronic patient records through health information exchanges.

Providers and patients value face time with each other. During EHR implementation, providers might spend less time in communication with patients. Providers must adapt their processes and clinic-to-administrative schedules. Any disruption or action that is perceived as deleterious to this relationship could result in a negative reaction to EHR implementation. As a result, physicians might oppose EHR adoption, or they might simply support the EHR solution with the shortest implementation time or least administrative burden. Patients might not like the reduced face time with the provider, but they might be attracted to EHR components such as e-prescribing, e-results, personal health records, and email access to the provider. These desirable features are available to the patient when the health care organization chooses to adopt various portions of the CPOE component to the EHR.

Results

Chosen Articles

The articles chosen for final inclusion were read once more to make a list of variables. The variables from the studies were listed as internal and external. There were significant commonalities in the variables used, so they were combined in the model.

EHR Adoption and Internal Organizational Influence

As depicted in Figure 1, 16 references identified internal factors [7,8,12,14,15,19,21-24,27-33]. Six identified size of the health care organization, and six identified strategic alliances. Five identified ownership and five identified complexity of care. Four identified capital expenditures. Three identified users, and three identified teaching status. Two identified user attitude toward HIS, and two identified communication among users. Workflow impact, interoperability, technical support, expert support, physician arrangements, unity of effort, and user computer anxiety were all identified by one study, independently.

The dependent variable was not consistent: seven references used EHR adoption [23-25,28,30,31,33], two used "electronic capture of clinical data," [23,25] one used a generic DV of technology adoption [15], and six used CPOE [20,23,25,28,29,31].

EHR Adoption and External Environmental Influence

As depicted in Figure 1, nine studies identified external environmental factors [12,14,15,19,22,24,28,30,31]. Five studies identified buyers, four studies identified patients, three studies identified competitiveness, two studies identified location, and one identified interdependence factors external to the organization that are associated with adoption of the EHR.

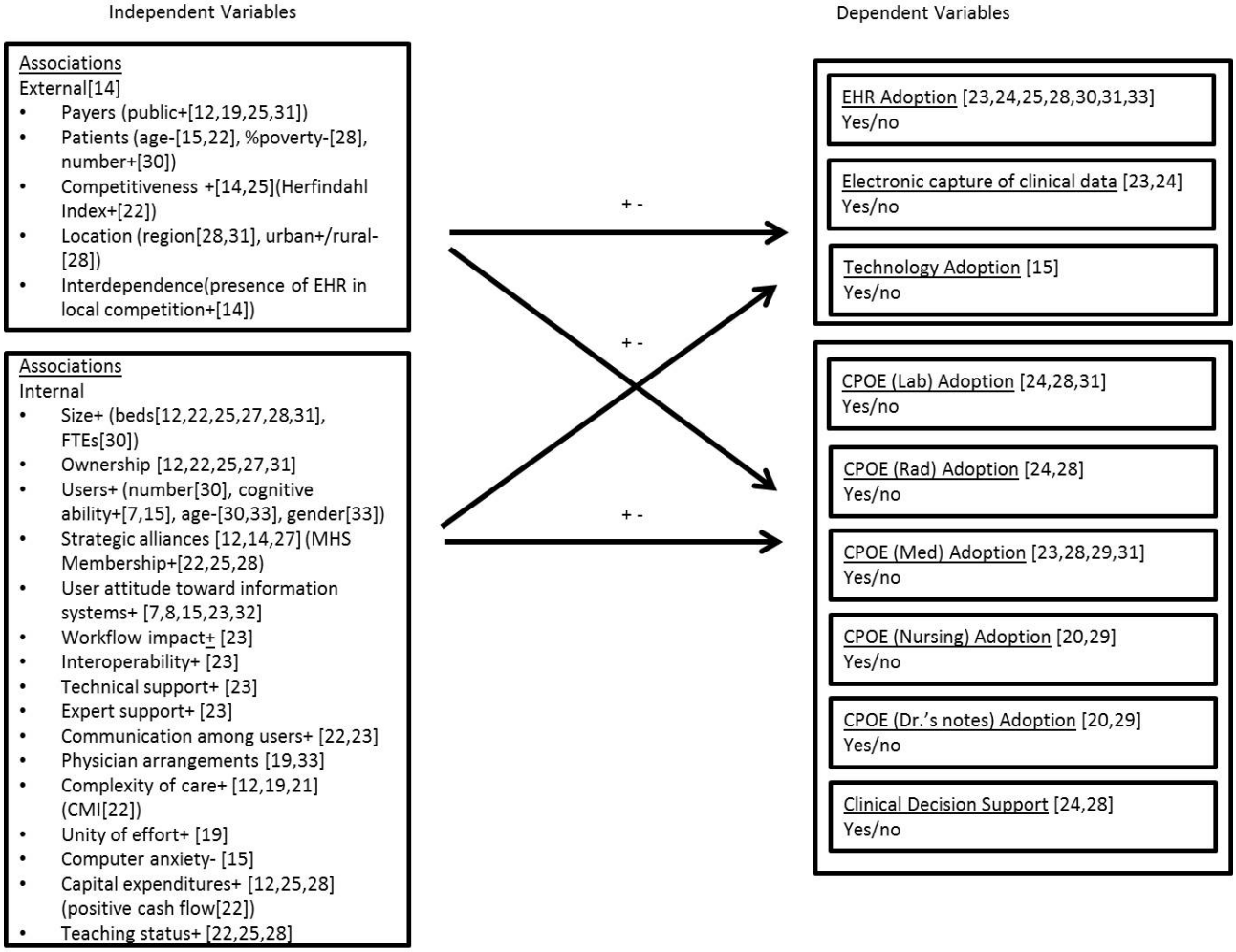
Overview of the Conceptual Model

As previously stated, there was overlap between the sources/theories. There were four internal forces and seven external forces identified through multiple works by three authors [11-14,22]. However, it was unclear in existing literature the degree to which these forces can influence a health care organization's decision to adopt the EHR. A complex conceptual model should provide insight into the strength and direction of the influence on the complex health care organization. The resulting conceptual model, depicted in Figure 2, posits a complex relationship between environmental influences, organizational strategy, and EHR adoption.

This framework captures both internal and external factors that influence the adoption of the EHR. The positive (+) and negative (−) signs in the model describe the relationship identified by the associated authors. For instance, Gin et al [24] identified a positive relationship between the external environmental factors of public payer (IV) and competitiveness (IV) and an association with the adoption of an EHR (DV). That is to say, the greater the percentages of an organization's reimbursements that come from a public source like CMS, the stronger the association of the organization's adoption of the EHR. Likewise, the greater the Herfindahl Index of the local competitive environment, the

stronger the association of the organization's adoption of the EHR. Age is another interesting factor because of its negative relationship with adoption. The older the patient population (external environmental IV) [15] and provider population (internal organizational IV) [30], the lower association with the adoption of the EHR (DV). The + and − signs above the arrows between the IVs and DVs indicates the variety of positive and negative associations with the adoption of an EHR.

Figure 2. Conceptual model of factors associated with adoption of the EHR.



Discussion

Principal Findings

The main findings of this study were that nine studies identified external factors and 16 studies identified internal factors associated with the adoption of EHR. These factors were depicted in a conceptual model to describe relationships to EHR adoption.

The conceptual model for EHR adoption illustrates a framework within which both administrators and policy makers can work to understand the levers that exert significant influence in the adoption of EHR. The extensive literature review conducted by this study builds a unique model from which empirical studies can be designed.

Identifying relationships between the adoption factors and adoption of the EHR becomes significant because it identifies levers that will produce a desired action. For instance, if a hospital has a majority of senior providers, perhaps from the Baby Boom generation, the administrators become aware of the additional effort that needs to go into user acceptance. A hospital that has a majority of new providers will not need to expend the resources on user acceptance, because studies already show a penchant for technology in younger generations. Similar inferences on other factors of adoption could be made, and some would require additional study.

For instance, the literature on workflow impact is split. There seems to be evidence that the presence of the EHR both enhances and complicates the providers' workflow. This observation clearly begs additional questions. Were subjects for the data in different phases of adoption of the EHR? Was the hospital that responded negatively in the middle of an EHR implementation? Logically, a large implementation of any technology will become disruptive to the organization. Several studies could emerge from this relationship alone.

Empirical models could easily be designed to further investigate specific relationships between the IVs and DVs. The set of studies on CPOE was interesting. Although there were some overlaps with adoption of the EHR, there were also studies that only looked at CPOE. There does not seem to be an abundance

of evidence in the literature about CPOE, and yet the AHA regularly collects data on six different versions of CPOE: laboratory, radiology, pharmacy, nursing, physician notes, and consults. There was no data to be found about the use or efficacy on CPOE consults. It might be interesting to determine the reason for this paucity of data.

Limitations

The EHR adoption conceptual model associates internal and external factors with the adoption of the EHR, but it is primarily based on an extensive literature review. So far, it is not empirically tested. However, data are available to test the theory. Because the findings of our study are descriptive in nature, we do not opine on appropriate medical use of the information.

Caution should be identified with the interaction of variables. Some variables will most likely confound or mask the effects of others. For instance, is there a direct relationship between the number of beds of a hospital and the number of full-time equivalents? There are staffing models that would most likely answer that question. If there is a similar relationship, then one of these variables should be eliminated in favor of the stronger one. Otherwise, the effects of the weaker variable will be masked by the other. A false conclusion could easily be identified concerning the masked variable.

A majority of references for this study were from the United States, with one exception from Hong Kong. The internal validity of this study is strongest within the US health care sector. The conceptual model might be limited outside the United States because of the nature of competition between hospitals. The analysis of 83 articles identified studies that used similar methods: survey or secondary data analysis. Many authors analyzed data from the AHA, a well-established data set in the United States. These data are self-reported, which comes with limited bias.

Conclusions

This study also identified overlap between studies in terms of variables. One interpretation of this overlap could be that the variables and associated studies are highly reliable. The key word/phrase searches described in the Methods section identifies the databases queried and results given. Other researchers should be able to duplicate or update this conceptual model going forward.

Authors' Contributions

The topic of research and conceptual model in this original article are from a chapter of the PhD dissertation of Kruse C DeShazo J, and Kim F sat on the dissertation committee, which helped direct the research and develop the model. Fulton L helped Kruse C extract portions of the dissertation worthy of publication.

Conflicts of Interest

None declared.

References

1. Health Information Technology for EconomicClinical Health A, 42 U. C. S URL: http://www.healthit.gov/sites/default/files/hitech_act_excerpt_from_arra_with_index.pdf [accessed 2014-03-18] [WebCite Cache ID 6OApfh2QK]
2. Elnahal SM, Joynt KE, Bristol SJ, Jha AK. Electronic health record functions differ between best and worst hospitals. Am J Manag Care 2011;17(4):e121-e147 [FREE Full text] [Medline: 21774097]
3. Song PH, McAlearney AS, Robbins J, McCullough JS. Exploring the business case for ambulatory electronic health record system adoption. J Healthc Manag 2011;56(3):169-180; discussion 181. [Medline: 21714372]
4. Institute of Medicine. Key Capabilities of an Electronic Health Record. 2003. URL: http://www.iom.edu/Reports/2003/Key-Capabilities-of-an-Electronic-Health-Record-System.aspx [accessed 2014-03-20] [WebCite Cache ID 6ODtL6HKG]
5. 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;24(5):1103-1117 [FREE Full text] [doi: 10.1377/hlthaff.24.5.1103] [Medline: 16162551]
6. Davis F. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989;13(3):319-340. [doi: 10.2307/249008]
7. Venkatesh V, Davis FD. A theoretical extension of the Technology Acceptance Model: four longitudinal field studies. Management Science 2000;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
8. Venkatesh V, Morris M, Davis G, Davis F. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425-478. [doi: 10.2307/30036540]
9. Aldrich HE, Pfeffer J. Environments of organizations. Annu Rev Sociol 1976;2(1):79-105. [doi: 10.1146/annurev.so.02.080176.000455]
10. Pfeffer J, Salancik G. The External Control of Organizations: A Resource Dependence Perspective. New York, NY: Harper & Row; 1978.
11. Scholz CE. Applying Rogers' Theory of Diffusion of Innovations to Examine Older Females' Perceptions of Size Labels For Apparel. Ottawa, Canada: National Library of Canada; 2002.
12. Rogers E. Diffusion of Innovations, 4th Ed. New York, NY: Free Press; 1995.
13. Rogers EM. Diffusion of innovations. New York, NY: Free Press; 2003.

14. Pfeffer J, Salancik G. The External Control of Organizations a Resource Dependence Perspective. Stanford, CA: Stanford Business Books; 2003.
15. Czaja SJ, Charness N, Fisk AD, Hertzog C, Nair SN, Rogers WA, et al. Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). Psychol Aging 2006;21(2):333-352 [FREE Full text] [doi: 10.1037/0882-7974.21.2.333] [Medline: 16768579]
16. Fumasoli T, Lepori B. Patterns of strategies in Swiss higher education institutions. In: High Educ. Netherlands: Springer; 2010:157-178.
17. Henderson J. Health Economics and Policy (with Economic Applications), 2nd ed. Mason, OH: South-Western College Pub; 2012.
18. Sikka V, Luke RD, Ozcan YA. The efficiency of hospital-based clusters: evaluating system performance using data envelopment analysis. Health Care Manage Rev 2009;34(3):251-261. [doi: 10.1097/HMR.0b013e3181a16ba7] [Medline: 19625830]
19. Bazzoli GJ, Shortell SM, Dubbs N, Chan C, Kralovec P. A taxonomy of health networks and systems: bringing order out of chaos. Health Serv Res 1999;33(6):1683-1717 [FREE Full text] [Medline: 10029504]
20. Ash JS, Bates DW. Factors and forces affecting EHR system adoption: report of a 2004 ACMI discussion. J Am Med Inform Assoc 2005;12(1):8-12 [FREE Full text] [doi: 10.1197/jamia.M1684] [Medline: 15492027]
21. Wang BB, Wan TT, Burke DE, Bazzoli GJ, Lin BY. Factors influencing health information system adoption in American hospitals. Health Care Manage Rev 2005;30(1):44-51. [Medline: 15773253]
22. Castillo VH, Martínez-García AI, Pulido JR. A knowledge-based taxonomy of critical factors for adopting electronic health record systems by physicians: a systematic literature review. BMC Med Inform Decis Mak 2010;10:60 [FREE Full text] [doi: 10.1186/1472-6947-10-60] [Medline: 20950458]
23. Blavin FE, Buntin MJ, Friedman CP. Alternative measures of electronic health record adoption among hospitals. Am J Manag Care 2010;16(12 Suppl HIT):e293-e301. [Medline: 21322296]
24. Ginn GO, Shen JJ, Moseley CB. Hospital financial position and the adoption of electronic health records. J Healthc Manag 2011;56(5):337-350; discussion 351. [Medline: 21991681]
25. Farley DE, Hogan C. Case-mix specialization in the market for hospital services. Health Serv Res 1990;25(5):757-783 [FREE Full text] [Medline: 2123838]
26. Centers for Medicare and Medicaid Services. NHE Fact Sheet. 2014. URL: http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet.html [accessed 2014-03-18] [WebCite Cache ID 6OAbW6btM]
27. Bazzoli GJ, Chan B, Shortell SM, D'Aunno T. The financial performance of hospitals belonging to health networks and systems. Inquiry 2000;37(3):234-252. [Medline: 11111282]
28. DesRoches CM, Campbell EG, Vogeli C, Zheng J, Rao SR, Shields AE, et al. Electronic health records' limited successes suggest more targeted uses. Health Aff (Millwood) 2010;29(4):639-646 [FREE Full text] [doi: 10.1377/hlthaff.2009.1086] [Medline: 20368593]
29. Harshberger CA, Harper AJ, Carro GW, Spath WE, Hui WC, Lawton JM, et al. Outcomes of computerized physician order entry in an electronic health record after implementation in an outpatient oncology setting. J Oncol Pract 2011;7(4):233-237 [FREE Full text] [doi: 10.1200/JOP.2011.000261] [Medline: 22043187]
30. Decker SL, Jamoom EW, Sisk JE. Physicians in nonprimary care and small practices and those age 55 and older lag in adopting electronic health record systems. Health Aff (Millwood) 2012;31(5):1108-1114. [doi: 10.1377/hlthaff.2011.1121] [Medline: 22535502]
31. Wolf L, Harvell J, Jha AK. Hospitals ineligible for federal meaningful-use incentives have dismally low rates of adoption of electronic health records. Health Aff (Millwood) 2012;31(3):505-513. [doi: 10.1377/hlthaff.2011.0351] [Medline: 22392661]
32. Lluch M. Healthcare professionals' organisational barriers to health information technologies-a literature review. Int J Med Inform 2011;80(12):849-862. [doi: 10.1016/j.ijmedinf.2011.09.005] [Medline: 22000677]
33. Cheung C, Tong E, Cheung N, Chan W, Wang H, Kwan M, et al. Factors associated with adoption of the electronic health record system among primary care physicians. JMIR Med Inform 2013;1(1):e1. [doi: 10.2196/medinform.2766]

Abbreviations

AHA: American Hospital Association
ARRA: American Recovery and Reinvestment Act
CMS: Center for Medicare and Medicaid Services
CPOE: computerized provider order entry
DV: dependent variables
EHR: electronic health record
EMR: electronic medical record
HHS: the Department of Health and Human Services

HIMSS: Health Information Management Systems Society
HIS: health information system
HIT: health information technology
HITECH: Health Information Technology for Economic and Clinical Health
IT: information technology
IV: independent variables
LR: long run
MHS: multihospital system
SR: short run
TAM: Technology Acceptance Model

Edited by G Eysenbach; submitted 19.11.13; peer-reviewed by V Castillo, S Speedie; comments to author 14.03.14; revised version received 03.04.14; accepted 28.04.14; published 23.05.14.

Please cite as:
Kruse CS, DeShazo J, Kim F, Fulton L
Factors Associated With Adoption of Health Information Technology: A Conceptual Model Based on a Systematic Review
JMIR Med Inform 2014;2(1):e9
URL: http://medinform.jmir.org/2014/1/e9/
doi:10.2196/medinform.3106
PMID:25599673

Original Paper

Automatically Recognizing Medication and Adverse Event Information From Food and Drug Administration's Adverse Event Reporting System Narratives

Balaji Polepalli Ramesh[1,2], MS Comp Sc, PhD; Steven M Belknap[3], MD; Zuofeng Li[1,2], MD, PhD; Nadya Frid[1,2], MS; Dennis P West[4], PhD; Hong Yu[1,5,6], PhD

[1]University of Massachusetts Medical School, Worcester, MA, United States

[2]University of Wisconsin Milwaukee, Milwaukee, WI, United States

[3]Northwestern University Feinberg School of Medicine, Department of Dermatology and Department of Medicine, Division of General Internal Medicine and Geriatrics, Chicago, IL, United States

[4]Northwestern University Feinberg School of Medicine, Department of Dermatology, Chicago, IL, United States

[5]VA Central Western Massachusetts, Leeds, MA, United States

[6]University of Massachusetts, Department of Computer Science, Amherst, MA, United States

Corresponding Author:
Hong Yu, PhD
University of Massachusetts Medical School
AS6-2071
368 Plantation St
Worcester, MA, 01605
United States
Phone: 1 (508) 856 3474
Fax: 1 (508) 856 8993
Email: Hong.Yu@umassmed.edu

Abstract

Background: The Food and Drug Administration's (FDA) Adverse Event Reporting System (FAERS) is a repository of spontaneously-reported adverse drug events (ADEs) for FDA-approved prescription drugs. FAERS reports include both structured reports and unstructured narratives. The narratives often include essential information for evaluation of the severity, causality, and description of ADEs that are not present in the structured data. The timely identification of unknown toxicities of prescription drugs is an important, unsolved problem.

Objective: The objective of this study was to develop an annotated corpus of FAERS narratives and biomedical named entity tagger to automatically identify ADE related information in the FAERS narratives.

Methods: We developed an annotation guideline and annotate medication information and adverse event related entities on 122 FAERS narratives comprising approximately 23,000 word tokens. A named entity tagger using supervised machine learning approaches was built for detecting medication information and adverse event entities using various categories of features.

Results: The annotated corpus had an agreement of over .9 Cohen's kappa for medication and adverse event entities. The best performing tagger achieves an overall performance of 0.73 F1 score for detection of medication, adverse event and other named entities.

Conclusions: In this study, we developed an annotated corpus of FAERS narratives and machine learning based models for automatically extracting medication and adverse event information from the FAERS narratives. Our study is an important step towards enriching the FAERS data for postmarketing pharmacovigilance.

(JMIR Med Inform 2014;2(1):e10) doi:10.2196/medinform.3022

KEYWORDS

natural language processing; pharmacovigilance; adverse drug events

Introduction

Background

An adverse event (AE) is an injury or untoward medical occurrence to a patient or clinical investigation subject who has been administered a pharmaceutical product and the AE does not necessarily have a causal relationship with the administered treatment [1,2]. An adverse drug event (ADE) is an injury resulting from a medical intervention related to a drug, including harm caused by the drug (adverse drug reactions and overdoses), and harm from the use of the drug (including dose reductions and discontinuations of drug therapy) [3,4]. Studies have reported that ADEs account for nearly 20% of all adverse events that occur in hospitalized patients [5-7]. In the United States alone, ADEs account for more than 770,000 injuries and deaths annually [8-10], and an increased average length of stay in hospitals at a cost of between $1.56 and $5.60 billion annually [3,11]. Improved methods for ADE detection and analysis may identify novel drug safety signals and lead to improved methods for avoiding ADEs, with their attendant burden of morbidity, mortality, and cost. As part of a major effort to support postmarketing drug safety surveillance, the US Food and Drug Administration (FDA) receives mandatory reports on ADEs from manufacturers through the FDA Adverse Event Reporting System (FAERS). The FAERS is a database that captures information concerning adverse events and medication errors associated with FDA-approved prescription drugs. Currently, FAERS contains over four million reports of adverse events dating from 1969 to present [12]. It serves as a rich resource for pharmacovigilance-the study of drug-related injuries for the purpose of making warning or withdrawal recommendations for pharmaceutical products [4]. A typical FAERS report incorporates both structured data and unstructured free text, as shown in Figure 1. The structured data entries incorporate each patient's personal and demographic information, a list of prescribed drugs, and the class of drug reaction (in this example, "anaphylactic reaction") (Figure 1). The Event/Problem narrative contains additional information relevant to describing the event, assessing causality, and grading severity (Figure 1). In this example, the narrative text contains the phrase that indicates causality between paclitaxel and the anaphylactic reaction while "experienced a life threatening anaphylactic reaction" shows the severity of the event, which is not coded in the structured data.

Although FAERS is an excellent resource to study drug effects, as stated in Tatonetti et al [13], the structured data does not incorporate confounding factors including concomitant medications and patient medical histories, which limits FAERS' effectiveness for pharmacovigilance. In contrast, such confounding factors are frequently described in the FAERS narratives. Making this data computationally available is critical for pharmacovigilance.

Currently, manual abstraction is required for identification of relevant data in FAERS narratives. Manual abstraction is expensive and often not practical, given the current size of the FAERS dataset, which contains millions of records. Therefore, it is important to develop computational approaches to automatically extract information from FAERS narratives. In this study, we report the development of both a corpus of FAERS narratives annotated with medication and adverse event information and a Natural Language Processing (NLP) system called AETagger that automatically extracts this information from the narratives and is adapted from existing tools. This is an important step towards enriching the existing FAERS' capacity for pharmacovigilance.

Figure 1. A sample AERS Report with structured data and narrative text.

FDA Adverse Event Reporting System (FAERS)
ISR Report for report # ISR 4901614-2

Reporter Name: **Reporter Phone:**
Reporter Org: **Reporter City:**
Reporter Street: **Reporter State:**
Reporter Zip: **Reporter Country:**

Product(s)	**Route**	**Indication(s)**
TAXOL	INTRAVENOUS	BREASTCANCER
KYTRIL	INTRAVENOUS	PREMEDICATION
DECADRON PHOSPHATE	INTRAVENOUS	PREMEDICATION
GLUCOPHAGE		
ACIPHEX		
LEVAQUIN		
PYRIDIUM		

Reaction(s)
ANAPHYLACTIC REACTION
Disease/Surgical Procedure
FOOD ALLERGY
DRUG HYPERSENSITIVITY
DIABETES MELLITUS
BREAST CANCER
DEPRESSION

Event/ Problem Narrative

A nurse reported that a 60- year old female consumer experienced a delayed allergic reaction while on paclitaxel therapy. The patient received her first dose of paclitaxel 280 mg on [DATE]. On [DATE], hours later, the consumer presented at the emergency room with pain in her hands and feet, feeling of throat closing and a rash (generalized). Epinephrine was administered and her symptoms resolved. She was discharged to home on the same day. The patient's medical history was a significant for diabetes, left breast cancer, depression and allergiies of shell fish and sulfa. Supplemental information was received from the reporting nurse on [DATE]; The patient experienced a life threatening anaphylactic reaction after receiving one dose of paclitaxel, on [DATE] for treatment of breast cancer. The Anaphylactic reaction was characterized by symptoms of pain in her hands and feet, feeling of throat closing and rash (generalized).

Related Work

There is extensive research related to AE and ADE detection and analysis from a variety of data sources. Earlier work examined patients' paper medical records determining whether AEs and ADEs can be reliably abstracted based on the information conveyed in those records. For example, Hiatt et al (1989) [14] was among one of the early studies that defined an AE as an injury caused at least in part by medical mismanagement (negligence). They then manually abstracted ADEs from patients' paper-based clinical medical records. Similarly, other early studies (eg, [3,7,15]) defined AEs and ADEs and manually abstracted them from clinical records. These studies indicate the feasibility and value of clinical records for ADE surveillance and prevention.

When electronic medical records (EMRs) became available, computational approaches were developed to automatically identify AE and ADE information from EMRs. Studies used rule-based approaches for detecting ADEs from EMR data [16-18]. Tinoco et al [19] compared a rule-based computer surveillance system called Health Evaluation through Logical Processing (HELP) [20] with manual chart reviews on 2137 patient admissions. They reported that HELP detected as many ADEs as were found by manual chart review, suggesting that NLP systems could improve ADE detection from EMR narrative data.

Many studies applied NLP to detect AEs and then inferred a causality relationship between a drug and an AE (called an ADE) using logical rules, statistical analyses, and supervised machine learning (ML) approaches. Hazlehurst et al [21] developed MediClass, a knowledge-based system that deploys a set of domain-specific logical rules to medical concepts that are automatically identified from EMR narratives (eg, progress notes) or precoded data elements (eg, medication orders). The system achieved a precision of 64% for detecting vaccine-related AEs [22]. A number of studies applied the NLP system [23-25], MedLEE [26], to detect AEs from discharge summaries and hospitalization records. For example, Wang et al [23] applied MedLEE to detect terms and mapped them to the Unified Medical Language System (UMLS) semantic types. Subsequently, they detected medication and AEs when the terms were mapped to the UMLS concepts with the semantic types

of Clinical Drug (T200) and Disease or Symptom (T047), respectively. The causality relationship between a medication and an AE was extracted from 25K discharge summaries based on a χ^2-statistical analysis of medication and AE. Evaluation of seven drugs for known ADEs led to a recall and precision of 75% and 31% respectively. Aramaki et al [27] manually annotated 435 discharge summaries for drugs and ADEs and then applied supervised machine learning techniques to detect these named entities. They identified the causality between drugs and AEs using pattern matching and SVM techniques. They reported a recall and precision score of 0.81 and 0.87 for drug, and 0.80 and 0.86 for AE detection respectively. For inferring causality they achieved recall and precision of 0.92 and 0.41 using pattern matching, and 0.62 and 0.58 using SVM technique respectively.

In addition to EMRs, studies have explored other data sources for ADE information, including biomedical literature [28,29], social media and the Internet [30-32]. Shetty and Dalal [33] mined ADEs from PubMed citations. They first built a document classifier to identify relevant documents that incorporate ADE relationships using Medical Subject Headings (MeSH) terms. For example, if an article is assigned "chemically induced" or "adverse effects," then the article is likely to incorporate an ADE. They then identified ADE signals using disproportionality analysis in which the rate at which a particular AE of interest co-occurs with a given drug is compared to the rate an AE occurs without the drug in the collection. Their evaluation on a predefined set of 38 drugs and 55 AEs showed that their literature-based approach could uncover 54% of ADEs prior to FDA warnings.

There is a rich store of literature for ADE detection on Spontaneous Reporting Systems (SRS) such as the FAERS reports and WHO VigiBase [34]. Studies have explored several statistical data mining and machine learning techniques on SRS for the detection of ADE signals [13,35-60]. However, all aforementioned approaches for ADE detection from FAERS are based on its structured data. In this study, we report the development and evaluation of supervised machine learning approaches for automatically detecting medication information and adverse events from the FAERS narratives. We speculate that such information can be a useful addition to the FAERS structured data for ADE detection.

Methods

Annotation Data and Procedure

Through our collaboration at Northwestern University [61], we obtained a collection of 150 de-identified FAERS narratives; a sample is shown in Figure 1. The data collection originally came as a scanned PDF image file. With Institutional Review Board (IRB) approval from Northwestern University and University of Wisconsin Milwaukee, we manually transcribed the PDF file into a computer-readable text file.

We randomly selected a set of 28 narratives for developing the annotation guideline (Multimedia Appendix 1). Our annotation guideline was based on the i2b2 challenges in NLP for Clinical Data Medication Extraction [62,63]. A balanced interdisciplinary team consisting of a linguist (NF), a physician (SB), two informaticians (BPR and HY) and a physician informatician (ZFL) developed the annotation guideline through an iterative process. At the end of reviewing 28 narratives, we obtained a guideline that all the members of the team agreed upon.

Following the final annotation guideline, two annotators (ZFL, designated as AnnPhy, and NF, designated as AnnLing), both of whom were the primary annotators for the i2b2 medication event detection challenge [63] in which we participated, independently annotated the remaining 122 AERS narratives. The different backgrounds of the annotators aids in building a corpus that is both linguistically driven and clinically correct. A physician (SB) served as a tiebreaker and resolved annotation disagreements. This collection of 122 narratives is comprised of approximately 23,000 word tokens and the average number of words per narrative is 190.2 (SD 130.3).

The annotation was carried out using Knowtator [64], a plugin for Protégé [65]. The Knowtator interface allows users to define entities that need to be annotated and configure the relationships between them. The 122 annotated narratives were used as both training and testing data for machine learning approaches described below. The annotated data was grouped into four collections each containing 122 narratives: *AnnPhy* and *AnnLing* –data annotated by annotators AnnPhy (ZFL) and AnnLing (NF), respectively; *Comb* –a joint set of annotations agreed upon by both AnnPhy and AnnLing, and *Tie* –a joint set of AnnPhy and AnnLing annotations where disagreements were resolved by the tiebreaker SB. We also report Cohen's kappa, a well-known statistic used to assess the degree of Inter-Annotator Agreement (IAA) between annotators [66]. We use these four sets of data to capture all named entities and build robust supervised machine learning classifiers to identify them.

Supervised Machine Learning

Machine Learning Techniques

Three supervised machine learning approaches were explored for automatically identifying medication information and adverse events: Naïve Bayes (NB), Support Vector Machines (SVMs) and Conditional Random Fields (CRFs) [67]. We built NB and SVM classifiers using Weka [68] and the CRF model was built using the A Biomedical Named Entity Recognizer (ABNER) toolkit [69]. NB is a simple model that assumes all attributes of the examples which are independent of each other given the context of the class. SVMs are a well-known statistical machine learning algorithm and have shown very good performance in many classification tasks [70,71]. CRFs have shown success in named entity recognition in the biomedical domain [69,72].

Learning Features

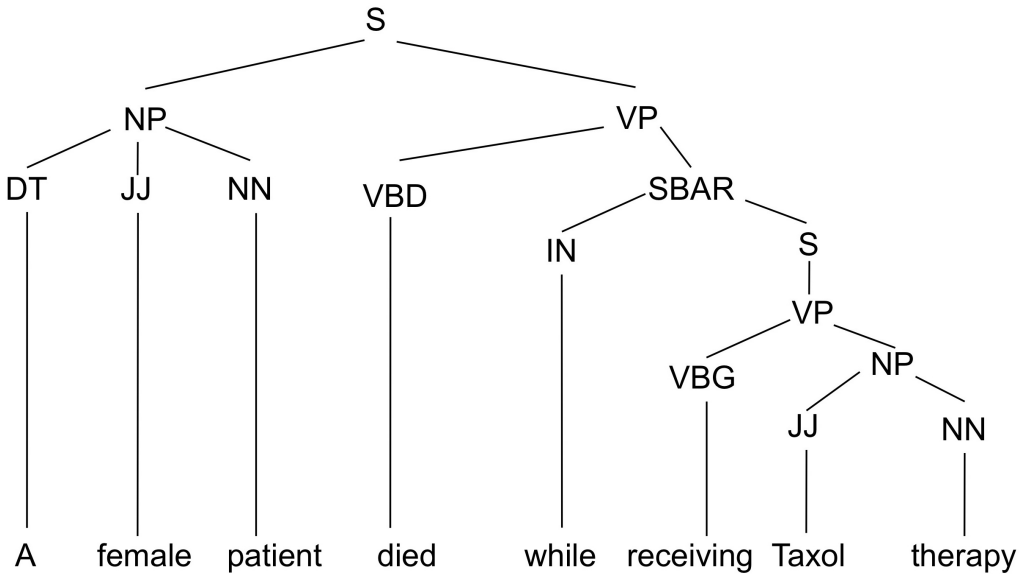
We explored a variety of features such as syntactic features, semantic features based on the external knowledge resource (UMLS), morphological and contextual features, presence of negation, hedging and discourse connectives as a feature in addition to ABNER default features which include bag of words and orthogonal features. We describe each of these in detail below.

The syntactic features include the part-of-speech (POS), the phrasal class of each token, and the POS of the token immediately to the left of the token under consideration. The syntactic features were extracted from the constituency parse tree generated by the Charniak-Johnson parser [73] trained in the biomedical domain. This parser was determined to have the best performance when tested on the GENIA corpus [74]. Figure 2 shows a sample constituency parse tree. In this example, the POS features determiner (DT), adjective (JJ), noun (NN) are the POS of tokens "A", "female", and "patient" respectively. Further, the phrasal class for all the three tokens is noun phrase (NP). The left sibling POS value of "A" is NONE assuming it is the start of the sentence. The left sibling POS of "female" and "patient" tokens are DT and JJ respectively.

We applied the UMLS Metamap [75,76] to extract semantic features, which are concepts and semantic types represented in the UMLS Metathesaurus. The morphological features were obtained by considering various characteristics of the word. We took attributes of the word, such as whether it was a digit, was capitalized, its alphanumeric order (ie, if the token started with letters and was followed by numerals or vice versa), and the presence of punctuation such as commas and hyphens. These features were extracted using a simple pattern-matching technique. The first (prefix) and last (suffix) three and four characters of the token were added as affix features.

We added as features, negation and hedging cues with their scope that were detected automatically by the systems described in the literature [77,78]. We also added presence of discourse connectives that were automatically detected by the discourse parser [79].

Figure 2. The sample constituency parse tree. S: simple declarative clause, NP: noun phrase, VP: verb phrase, DT: determiner, JJ: adjective, NN: noun, VBD: verb, past tense, SBAR: subordinate clause, IN: preposition or subordinating conjunction, VBG: verb, gerund or present participle.



Systems

Overview

We developed several taggers to evaluate the complexity of the task for identifying medication information and adverse events and the impact of features.

Systems to Evaluate Task Complexity

In this experiment, we built two baseline systems to compare the performance of ML algorithms. The first system *BaseDict* is a simple dictionary-matching system. A lexicon of medications and AEs is compiled from the UMLS Metathesaurus using the semantic types as defined by Wang et al [23], where terms having the semantic types Clinical Drug (T200) and Disease or Symptom (T047) were considered as drug and adverse event respectively. The baseline system *BaseDict*, tags all instances of the lexicon that match within the text. The second system, *MetaMapTagger*, is a UMLS Metamap [75] based system that tags phrases as AEs or medications using UMLS semantic types similar to *BaseDict*.

The baseline systems were compared with taggers built using bag of words as the default feature –*NBTagger*, a NB-based tagger, *SVMTagger*, a SVM-based tagger, and *SimpleTagger*, a CRF-based tagger built using ABNER default features. We then evaluate the taggers by adding all the features defined in the Learning Features section, which we call *NBTagger*$^+$, *SVMTagger*$^+$ and *CombinedTagger* for NB, SVM- and CRF-based taggers respectively.

Systems to Evaluate Impact of Features

We evaluate the impact of various features on the performance of tagger. We used the ML technique found to have the best performance in our previous experiment. In addition to the default features trained as *SimpleTagger*, we individually added syntactic features (*SyntacticTagger)*, semantic features (*SemanticTagger*), morphological features (*MorphologicalTagger*), affix features (*AffixTagger)*, negation and hedging features (*NegHedgeTagger*), discourse connective features (*ConnectiveTagger*), and a tagger incorporating all the features (*CombinedTagger*) which were trained to identify the named entities.

Machine Learning Evaluation Metrics

All the AE taggers trained were evaluated using ten-fold cross-validation. We reported recall, precision, and F1 score. Recall is the ratio of the number of entities of a certain class correctly identified by the system and the number of entities of that class in the gold standard. Precision is the ratio of the number of entities of a certain class correctly identified by the system and the number of entities of that class predicted by the system. F1 score is the harmonic mean of precision and recall.

Results

Corpus Characteristics and Annotation Agreement

Table 1 shows the definitions of adverse event and medication-related named entities, the number of annotated instances, and Cohen's kappa value. The annotation agreement is calculated based on two criteria: *strict* in which the two annotations have an exact match, and *unstrict* in which there exists an overlap of at least one word between the two annotations. We measured the agreement using *unstrict* criteria to estimate the agreement between annotators when entity boundary is ignored. The table also shows the number of instances annotated in all four data sets.

As shown in Table 1, *adverse event* (*AE*) was the most frequently annotated entity followed by *medication* entity. *Duration* had the least number of annotated instances and lowest kappa value (.34) for *strict* criteria. *Indication* had the second highest kappa value for *unstrict* criteria (.93) after *medication* (.95), since most of the *indication* entities were followed by explicit and unambiguous patterns such as "for the treatment of", "diagnosed with", "due to", "enrolled in breast cancer study", and so on.

Table 1. Named entity definition, number of annotated instances, and inter-annotator agreement measured by Cohen's kappa for both strict and unstrict criterion.

Named entity	Definition	Number of instances annotated				kappa (strict)	kappa (unstrict)
		AnnPhy	AnnLing	Comb	Tie		
Medication	Name of the drug they administered to patient including drug class name or medications referred to with	1231	1278	1152	1286	.92	.95
Dosage	Amount of a single medication used in each administration	143	315	137	205	.59	.82
Route	Method for administering the medication	115	244	107	132	.59	.64
Frequency	How often each dose of the medication should be taken	25	56	21	42	.58	.74
Duration	How long the medication is to be administered	34	153	24	51	.34	.87
Indication	Medical conditions for which the medication is given	175	148	126	175	.76	.93
Adverse event (AE)	Harm directly caused including the pronouns referring to it by the drug at normal doses and during normal use	1689	2083	1646	1842	.83	.93
Other signs, symptoms and diseases (OSSD)	Other symptoms associated with the disease	234	140	90	147	.50	.71
Treatment	Treatment the patient received for the disease	77	216	62	153	.39	.77
Total		3723	4633	3365	4033		

Results of Supervised Learning

Table 2 reports recall, precision, and F1 score of the AETaggers for identifying the AE and other medication-related named entities on each of the four data sets as described in Annotation and data procedure section.

The baseline system *BaseDict* that matches only *AE* and *medication* achieved an F1 score of 0.45, 0.41, 0.46, and 0.42 on the *AnnPhy*, *AnnLing*, *Comb*, and *Tie* datasets respectively. The *MetamapTagger* also had similar performance. Among the taggers using bag of words as features, the CRF-based *SimpleTagger* had the best performance. The addition of features improved the performance of the ML classifiers. The *CombinedTagger* achieved best performance with F1 scores of 0.69, 0.74, and 0.73 on the *AnnPhy*, *AnnLing*, and *Comb* datasets respectively. The *SVMTagger*$^+$ had the best performance with a 0.66 F1 score on the *Tie* dataset. The difference in performance between *CombinedTagger* and *SVMTagger*$^+$ taggers was statistically significant only on *AnnLing* dataset (*t* test, *P*=.003). The ML-based taggers clearly outperform the baseline method. The CRF-based tagger had the best overall performance and was therefore chosen as the system to be adopted for subsequent experiments measuring impact of features.

We trained the CRF-based AETaggers using different features as described in the Learning Features section. The results show that the *CombinedTagger* achieved the highest performance on all datasets. Our results also show that the *AnnLing* dataset has the highest performance while *Tie* performs the lowest. *Comb* outperforms both *Tie* and *AnnPhy*.

Since the *Comb* dataset's performance (0.73 F1 score) is close to the highest (0.74 F1 score) and contains annotations agreed upon by both annotators, we further report feature analyses using the *Comb* dataset. Table 3 shows how different learning features affect AETagger's performance. The results show that adding a single feature added little to the overall performance, although the performance of different entities varied. Affix features improved *route* and *duration* but decreased *AE*, *medication*, and *dosage*. Connective features increased the performance of *dosage*, *route*, and *indication*; however, the performance of *medication* decreased. Other features (morphological, negation, hedge, semantic, and syntactic) showed similar patterns. On the other hand, when all features were added, the overall performance increased to 0.73 F1 score (default 0.71), although the increase was not statistically significant (*t* test, *P*=.08).

Table 2. The precision, recall, and F1 score of Taggers on each of the four annotated data sets (*t* test, *P*<.01).

Machine learning	AnnPhy Mean (SD)			AnnLing Mean (SD)			Combined Mean (SD)			Tie Mean (SD)		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Task complexity												
BaseDict	0.45 (0.10)	0.86 (0.08)	0.31 (0.09)	0.41 (0.09)	0.91 (0.07)	0.27 (0.08)	0.46 (0.12)	0.82 (0.06)	0.32 (0.11)	0.42 (0.10)	0.86 (0.13)	0.28 (0.08)
MetaMapTagger	0.41 (0.17)	0.41 (0.16)	0.42 (0.18)	0.41 (0.10)	0.47 (0.20)	0.37 (0.15)	0.42 (0.18)	0.41 (0.17)	0.43 (0.19)	0.40 (0.16)	0.46 (0.19)	0.36 (0.14)
NBTagger	0.22 (0.08)	0.39 (0.17)	0.15 (0.05)	0.23 (0.08)	0.45 (0.14)	0.16 (0.06)	0.24 (0.08)	0.40 (0.17)	0.17 (0.05)	0.20 (0.06)	0.47 (0.19)	0.13 (0.04)
SVMTagger	0.55 (0.05)	0.77 (0.10)	0.44 (0.04)	0.55 (0.05)	0.78 (0.07)	0.43 (0.05)	0.58 (0.04)	0.78 (0.10)	0.46 (0.04)	0.59 (0.04)	0.80 (0.05)	0.46 (0.05)
SimpleTagger	0.67 (0.09)	0.77 (0.09)	0.60 (0.09)	0.72 (0.08)	0.81 (0.06)	0.66 (0.10)	0.71 (0.08)	0.81 (0.09)	0.63 (0.08)	0.63 (0.09)	0.69 (0.08)	0.55 (0.10)
NBTagger^{+}	0.45 (0.09)	0.38 (0.10)	0.56 (0.06)	0.44 (0.06)	0.39 (0.07)	0.50 (0.06)	0.46 (0.09)	0.37 (0.11)	0.60 (0.04)	0.43 (0.07)	0.38 (0.08)	0.51 (0.07)
SVMTagger^{+}	0.66 (0.07)	0.78 (0.10)	0.58 (0.06)	0.67 (0.07)	0.78 (0.07)	0.59 (0.07)	0.70 (0.06)	0.80 (0.11)	0.63 (0.05)	0.66 (0.07)	0.78 (0.06)	0.57 (0.08)
CombinedTagger	0.69 (0.09)	0.77 (0.10)	0.62 (0.09)	0.74 (0.08)*	0.81 (0.07)	0.68 (0.09)	0.73 (0.08)	0.81 (0.10)	0.66 (0.07)	0.65 (0.08)	0.71 (0.08)	0.60 (0.09)
Impact of features												
SimpleTagger	0.67 (0.09)	0.77 (0.09)	0.60 (0.09)	0.72 (0.08)	0.81 (0.06)	0.66 (0.10)	0.71 (0.08)	0.81 (0.09)	0.63 (0.08)	0.63 (0.09)	0.69 (0.08)	0.55 (0.10)
AffixTagger	0.67 (0.09)	0.78 (0.09)	0.60 (0.09)	0.73 (0.09)	0.81 (0.06)	0.66 (0.10)	0.70 (0.08)	0.81 (0.09)	0.63 (0.08)	0.61 (0.09)	0.70 (0.08)	0.52 (0.10)
ConnectiveTagger	0.67 (0.09)	0.77 (0.09)	0.60 (0.09)	0.73 (0.08)	0.81 (0.06)	0.66 (0.10)	0.71 (0.08)	0.81 (0.09)	0.63 (0.08)	0.63 (0.09)	0.70 (0.07)	0.57 (0.10)
MorphologicalTagger	0.68 (0.09)	0.77 (0.08)	0.60 (0.10)	0.73 (0.08)	0.81 (0.06)	0.66 (0.09)	0.71 (0.08)	0.80 (0.09)	0.63 (0.08)	0.64 (0.08)	0.71 (0.07)	0.59 (0.09)
NegHedgeTagger	0.66 (0.09)	0.77 (0.09)	0.59 (0.10)	0.72 (0.08)	0.81 (0.06)	0.65 (0.10)	0.71 (0.08)	0.81 (0.09)	0.63 (0.08)	0.61 (0.09)	0.69 (0.08)	0.54 (0.10)
SemanticTagger	0.68 (0.09)	0.77 (0.10)	0.61 (0.09)	0.70 (0.09)	0.78 (0.07)	0.64 (0.10)	0.72 (0.09)	0.80 (0.11)	0.65 (0.08)	0.63 (0.09)	0.69 (0.10)	0.58 (0.09)
SyntacticTagger	0.68 (0.09)	0.78 (0.09)	0.61 (0.10)	0.72 (0.08)	0.80 (0.06)	0.65 (0.09)	0.71 (0.08)	0.80 (0.09)	0.64 (0.08)	0.63 (0.08)	0.70 (0.08)	0.58 (0.09)
CombinedTagger	0.69 (0.09)	0.77 (0.10)	0.62 (0.09)	0.74 (0.08)	0.81 (0.07)	0.68 (0.09)	0.73 (0.08)	0.81 (0.10)	0.66 (0.07)	0.65 (0.08)	0.71 (0.08)	0.60 (0.09)

Table 3. The F1 score of different named entities with different features on *Comb* dataset.

Feature group	AE mean (SD)	Medication mean (SD)	Dosage mean (SD)	Frequency mean (SD)	Route mean (SD)	Duration mean (SD)	Indication mean (SD)	OSSD mean (SD)	Treatment mean (SD)	Overall mean (SD)
Default	0.70 (0.10)	0.82 (0.10)	0.59 (0.35)	0.57 (0.46)	0.36 (0.33)	0.20 (0.42)	0.57 (0.12)	0.44 (0.45)	0.60 (0.52)	0.71 (0.08)
Affix	0.69 (0.11)	0.81 (0.12)	0.58 (0.37)	0.59 (0.45)	0.55 (0.37)	0.40 (0.52)	0.57 (0.09)	0.51 (0.44)	0.60 (0.52)	0.70 (0.08)
Connective	0.70 (0.10)	0.81 (0.10)	0.69 (0.31)	0.57 (0.46)	0.44 (0.36)	0.20 (0.42)	0.60 (0.15)	0.44 (0.45)	0.60 (0.52)	0.71 (0.08)
Morphological	0.70 (0.10)	0.82 (0.10)	0.57 (0.35)	0.59 (0.45)	0.32 (0.32)	0.20 (0.42)	0.62 (0.12)	0.47 (0.43)	0.60 (0.52)	0.71 (0.08)
NegHedge	0.69 (0.10)	0.82 (0.10)	0.56 (0.36)	0.59 (0.45)	0.36 (0.33)	0.20 (0.42)	0.59 (0.11)	0.50 (0.43)	0.60 (0.52)	0.71 (0.08)
Semantic	0.71 (0.11)	0.82 (0.11)	0.56 (0.35)	0.65 (0.40)	0.34 (0.33)	0.30 (0.48)	0.64 (0.13)	0.43 (0.39)	0.60 (0.52)	0.72 (0.09)
Syntactic	0.70 (0.10)	0.81 (0.11)	0.61 (0.35)	0.59 (0.45)	0.32 (0.31)	0.34 (0.47)	0.58 (0.11)	0.44 (0.45)	0.60 (0.52)	0.71 (0.08)
All	0.72 (0.10)	0.83 (0.11)	0.61 (0.37)	0.59 (0.44)	0.32 (0.31)	0.34 (0.47)	0.65 (0.11)	0.55 (0.39)	0.60 (0.52)	0.73 (0.08)

Annotation Disagreements

Overview

We manually analyzed the annotation disagreements and found they can be organized into three main categories: (1) boundary inconsistencies –disagreement due to assignment of inconsistent boundaries to entities; (2) missed named entity annotations –disagreement where one of the annotators annotated an entity and the other annotator missed it; (3) inconsistent named entity annotations –disagreement due to inconsistent categorization of entities.

There were a total of 2955 disagreed token instances, of which 1591 (53.84%) were related to *AE* and *medication* named entities.

Boundary Inconsistencies

We found that inconsistencies related to boundary accounted for nearly 13.94% (412/2955) of disagreement. In all the examples in the article, the named entity instance is shown in italics and the named entity type is shown within the "[]".

In Textbox 1 Example 1, AnnLing annotated "three hour" as *duration* and "infusion" as *route*, AnnPhy annotated "three hour infusion" as *duration* only. This inconsistency exemplifies differences between the linguist and the physician. While the linguist can separate the linguistic differences between different named entities, we found that physicians (both ZFL and SB) frequently overlook the differences, which leads to inconsistent annotations.

Textbox 1. Examples.

Example 1: She received approximately less than two minutes of therapy with intravenous Taxol (paclitaxel), 280 mg in a *three hour* [duration] *infusion* [route] for phase IIID ovarian cancer, when the symptoms occurred.

Example 2: The patient then became *lightheaded* [AE], *collapsed* [AE], and was *unconscious* [AE].

Example 3: Investigator considers that *haematologic toxicity* [AE] of methotrexate could be increased by interaction with apranax (naproxene) and sintrom (acenocoumarol).

Example 4: On [words marked], the patient died, presumed to be a result of *cardiogenic shock* [AE]. Prior to death, the patient was noted for having an increase in troponin T level, and found to be more unresponsive.

Example 5: Moderate anaphylactoid symptom appeared after administration of docetaxel and recovered later. After the end of administration, convulsion appeared. Anti-convulsion agent could not be administered due to *allergy* [AE].

Example 6: …days after the last Vinorelbine intake patient was hospitalized due to *NCI/CTC grade 4 neutropenia* [AE] without *fever* [OSSD]…

Missed Named Entity Annotations

Missed named entity annotation was the major cause for disagreement. Among 2955 disagreed token instances, 2355 or approximately 79.69% belong to this category. Table 4 shows instances of *medication* that were annotated by one annotator and missed by other. Examples 1-5 (Table 4) were annotated by AnnPhy but missed by AnnLing; examples 6-10 (Table 4) were annotated by AnnLing but missed by AnnPhy.

AnnLing explained that "blood transfusion", "fluids", and "red packed cells" shown in examples 1, 2, and 5 were not *medication*, but referred to a kind of treatment or medical procedure. In example 3, AnnLing missed annotating "normal saline" as *medication*. In example 4, "oxygen" was not annotated because AnnLing felt it did not represent *medication.* Annotators did not reach any consensus on annotating "oxygen" as *medication* or not. The differences here exemplify the strength of the physician as a domain-expert who may interpret the semantics of EMR notes more accurately than the linguist.

In examples 7, 8, and 10, in Table 4, AnnPhy did not annotate "treatment", "Re-exposure", and "chemotherapy" as these entities were anaphoric references; AnnLing, being a linguist, annotated these anaphoric references as *medication*. In example 6 (Table 4), AnnLing annotated "drug" as *medication* but AnnPhy did not annotate the entity because the text did not refer to any *medication*. Later, AnnLing agreed that where there is mention of entities but they do not refer to specific entities, such as "drug" in example 6, they should not be annotated. Example

9 in Table 4 was a special case where "concomitant drug" refers to the role or function of the drug, "Solupred", rather than referring to a drug. AnnPhy did not annotate such instances. These examples demonstrated that annotating medical texts is a complex and cumbersome task. Further refinement of guidelines in such instances may improve the consistency of annotations.

Table 4. Disagreement in medication annotation (medication text is italicized).

Annotation	Medication annotation
Annotated by AnnPhy but not annotated by AnnLing	1. Given multiple *blood transfusions* (hemoglobin: 4.8).
	2. Pressors continued with *fluids*.
	3. He was admitted to the hospital and hydrated with *normal saline*.
	4. The event was treated with steroids and *oxygen*.
	5. Pancytopenia, treated with G-CSF, erythropoetin, and *red packed cells*.
Annotated by AnnLing but not annotated by AnnPhy	6. Causality assessment *drug* relationship is unable to determine for Taxol.
	7. The 4th previous courses of *treatment* were well tolerated. 8. During the first infusion of paclitaxel, the patient experienced a decrease in blood pressure and was unconscious for a short while**.***Re-exposure* elicited the same symptoms.
	9. The *concomitant drug* prescribed was oral Solupred instead of Solumedrol.
	10. A female patient possibly received non-therapeutic dosages of intravenous Taxol (paclitaxel), Paraplatin (carboplatin), and/or Platinol (cisplatin) for the treatment of ovarian cancer and subsequently expired. It was reported that the pharmacist possibly diluted *the chemotherapy* improperly.

Inconsistent Named Entity Category Annotations

We have annotated a total of nine categories of named entities, as shown in Table 1. The third type of inconsistency was caused by inconsistent named entity assignments. Among 2955 disagreed token instances, 188 (6.36%) belong to inconsistent named entity categorization. We manually examined few instances and Examples 2-6 in Textbox 1 show the annotated sentences where inconsistency occurred. Example 2 in Textbox 1 is an example where both annotators agreed on the AE annotation.

Example 3 in Textbox 1, however, shows an instance where AnnPhy and the tiebreaker agreed on "haematologic toxicity" as an AE whereas AnnLing did not initially annotate the entity. This inconsistency suggests that domain knowledge is required for annotation. After discussion with two other annotators, AnnLing agreed that "haematologic toxicity" should be annotated an AE.

Example 4 in Textbox 1 shows an instance where AnnLing and the tiebreaker agreed on "cardiogenic shock" as an AE but AnnPhy annotated it as OSSD. AnnPhy argued that "cardiogenic shock" caused "death" and therefore "death" should be an AE and "cardiogenic shock" is the reason for death and therefore was annotated as OSSD. This example shows the complexity of clinical cause.

In Textbox 1 Example 5, the tiebreaker annotated "allergy" as an AE, whereas AnnPhy annotated it as OSSD and AnnLing did not annotate it as an AE because it refers to the patient's history of "allergy" and does not represent a current instance of AE. We will need to refine our annotation guideline to add current or past status in addition to the named entity annotation.

Example 6 in Textbox 1 shows an instance of boundary inconsistency. AnnPhy and AnnLing both annotated "NCI/CTC grade 4 neutropenia without fever" as an AE whereas the tiebreaker annotated "NCI/CTC grade 4 neutropenia" as an AE and "fever" as OSSD. This is a case in which annotators interpret clinical texts differently. Such an inconsistency is difficult to address due to the nature of ambiguity in clinical texts.

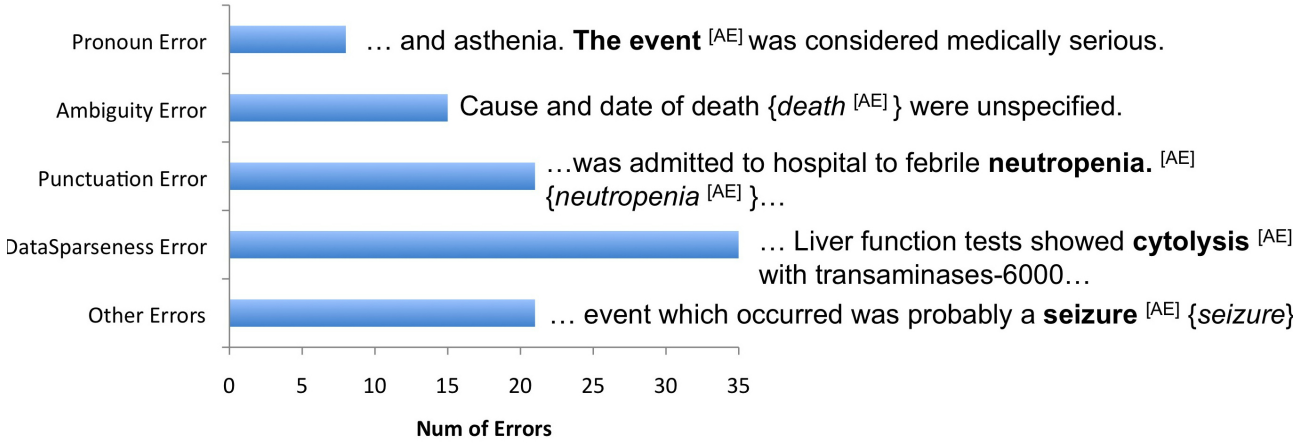
Error Analyses

For error analyses, we focused on *CombinedTagger* because it yielded the highest performance (as shown in Table 2) and the *Comb* dataset because it contained annotations agreed on by both annotators. We randomly selected 100 named entities predicted wrongly by *CombinedTagger* and manually analyzed them. As shown in Figure 3, we group the named entities into a total of five types of errors and give an illustrative example for each. In all examples, annotated named entities are shown in bold, the tagger output in {*italicized*} and the named entity type is shown within "[]". The leading type of error was data sparseness (35%). Data sparseness is a common problem and the major cause of poor performance. For instance, the gold standard consisted of a number of singleton instances (instances that appear only once) like "cytolysis", "sodium chloride solution 0.9% 100ml", and "neoplasm of unspecified nature of respiratory system" that created sparseness in the data.

The second cause of error was inconsistent inclusion of punctuation (21%). The gold standard had inconsistency in inclusion of punctuation (eg, a period [.] in "neutropenia.") as a part of a named entity. This boundary inconsistency reduced the overall performance. Figure 3 shows an instance where the gold standard included a period as part of named entity "neutropenia." but the tagger failed to include it ("neutropenia"). This was followed by an error caused by ambiguous named entities (15%). The instances in the gold standard that were assigned to multiple named entity categories resulted in ambiguous entities. For example, "death" was annotated as either *AE* or *OSSD*. This could have confused the ML algorithm and yielded a lower performance. In Figure 3, the instance "death" was not annotated as *AE* in the *Comb* dataset due to disagreements between annotators, but the tagger identified it as an *AE*. The missed pronoun annotations such as "the event" contributed to 8% of the errors. The final category was other

type of errors (21%), for which the exact cause of error could not be determined. In Figure 3, "seizure" was annotated as an *AE* but the tagger failed to identify it. The exact cause for miscategorization could not be determined.

Figure 3. Error categories, their frequency, and an illustrative example of error category on 100 randomly sampled instances. The annotated entities are shown in bold, the annotated named entity type is shown within "[]" and tagger output is {italicized}. AE: adverse events.



Annotation Inconsistencies

As predicted, annotation inconsistency played an important role on AETaggers' performance as our Pearson correlation results (coefficient of 0.73) show that the IAA value (Cohens' kappa) is positively correlated with the machine learning performance of named entity recognition. This is not surprising because inconsistent annotations confuse the machine learning systems.

Our manual analysis of inconsistency revealed that nearly 20% of errors were due to inconsistent inclusion of punctuation in annotation. When we removed the inconsistency in punctuation, the F1 score of *CombinedTagger* increased from 0.73 to 0.79, which was statistically significant (*t* test, *P*=.001). Although the missed pronoun annotations of *AE* and *medication* can be fixed readily, they also contributed to the lower performance of the tagger.

Data Sparseness

Data sparseness is a common problem and the major cause of poor performance. The performance of AETagger was positively correlated with the size of the annotated data for each named entity (a Pearson correlation coefficient of 0.64). In the cases of *frequency*, *duration*, *OSSD*, and *treatment* entities, data was very sparse (Table 1) and taggers showed low performance on these named entities. In addition to low performance, data sparseness also contributed to a higher standard deviation (Table 3). When the training data incorporate instances of a named entity but the testing data do not, the precision decreases. When the training data misses instances of a named entity but the testing data do not, then recall suffers.

Learning Features

To further understand the contribution of learning features on the performance of AETagger, we first trained the tagger with all the features and used it as a baseline system (*CombinedTagger*). We then removed each feature category one at a time. Table 5 shows the performance of taggers with each feature category removed. Consistent with Table 3, the results show that each feature contributed to the performance differently.

Table 5. The precision, recall, and F1 score of Taggers with feature categories removed one at a time on each of the four annotated data sets.

Tagger	AnnPhy Mean (SD)			AnnLing Mean (SD)			Combined Mean (SD)			Tie Mean (SD)		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
All features	0.67 (0.09)	0.77 (0.10)	0.62 (0.09)	0.74 (0.08)	0.81 (0.07)	0.68 (0.09)	0.73 (0.08)	0.81 (0.10)	0.66 (0.07)	0.65 (0.08)	0.71 (0.08)	0.60 (0.09)
No affix features	0.68 (0.09)	0.76 (0.10)	0.62 (0.09)	0.71 (0.10)	0.78 (0.07)	0.65 (0.11)	0.71 (0.09)	0.79 (0.11)	0.64 (0.09)	0.64 (0.08)	0.70 (0.08)	0.60 (0.08)
No connective features	0.69 (0.09)	0.77 (0.10)	0.62 (0.09)	0.74 (0.08)	0.81 (0.06)	0.69 (0.09)	0.73 (0.08)	0.81 (0.10)	0.66 (0.07)	0.65 (0.08)	0.71 (0.08)	0.60 (0.09)
No morphological features	0.69 (0.09)	0.78 (0.10)	0.62 (0.09)	0.73 (0.08)	0.81 (0.06)	0.66 (0.09)	0.73 (0.08)	0.82 (0.10)	0.66 (0.07)	0.65 (0.08)	0.72 (0.08)	0.60 (0.08)
No negation and hedge features	0.68 (0.09)	0.77 (0.10)	0.62 (0.09)	0.74 (0.08)	0.81 (0.07)	0.68 (0.09)	0.72 (0.08)	0.81 (0.10)	0.65 (0.07)	0.64 (0.09)	0.71 (0.09)	0.59 (0.09)
No semantic features	0.67 (0.08)	0.77 (0.08)	0.60 (0.09)	0.74 (0.08)	0.82 (0.05)	0.68 (0.10)	0.71 (0.08)	0.80 (0.09)	0.64 (0.08)	0.64 (0.08)	0.71 (0.07)	0.59 (0.08)
No syntactical features	0.68 (0.09)	0.77 (0.10)	0.61 (0.09)	0.73 (0.08)	0.80 (0.07)	0.68 (0.09)	0.71 (0.09)	0.80 (0.11)	0.64 (0.08)	0.64 (0.08)	0.70 (0.09)	0.58 (0.09)

Discussion

Principal Findings

Our results show that medication and adverse events can be reliably annotated (Cohen's kappa value of .64-.95 IAA as shown in Table 1) in the FAERS narratives. Many named entities (eg, *indication*) that had shown low annotation agreements in the i2b2 challenge [63] had good annotation agreements in our dataset. The improvements were attributed to improved annotation guidelines and the quality and domain specificities of the FAERS narratives.

With a good IAA, we still found room to further improve the annotation guideline. For example, our error analyses (Figure 3) show that inconsistencies were introduced by annotation boundary; therefore it can be further refined. Although *medication* had the highest IAA (.95), our analysis (Table 4) found that the inconsistency in *medication* was introduced by whether instances like "fluids" could be considered as medication or not. In the future, we may separate *medication* into two classes: *strict medication* and *unstrict medication*. The names and mentions of all drugs appearing in the United States Pharmacopeia will belong to *strict medication;* any substances or chemicals—including oxygen, fluids, drinks, and others–given to patients during the treatment will be classified as *unstrict medication.* Refining the guideline to annotate previous and potential AEs like "allergy" (Example 5) may further reduce the inconsistency.

We explored various ML methods and compared them with a baseline string matching and Metamap-based systems to assess the complexity of the task. The CRF-based tagger had the best performance. Further analyses of the CRF tagger found that data sparseness affected the taggers' performance (Figure 3). For example, the standard deviation of *treatment* is high because we found that the testing data did not incorporate *treatment* instances. Similar behavior was also observed for other sparse entities (Table 3).

Using the best performing ML technique, we explored a variety of features (Table 2 and Table 3). The features had a mixed effect on the performance of the taggers and the combination of all the features improved overall performance slightly. This suggests the robustness of the default features for CRFs. Since most of the features were extracted automatically (eg, negation, hedge cues, and discourse connectives were extracted using the taggers [77,78] and parser [79] we developed), the accuracy of the extracted features played an important role in overall performance of the tagger. To avoid the noise introduced by automatic feature extraction, one may explore the features manually annotated such as POS in the PennTree Bank [80]. This is, however, expensive. An alternative is to further improve the performance of the BioNLP systems for feature extraction.

Throughout the study, we found that additional features may be further included. For example, we observed that *OSSD* most often appeared in the patient's medical history. We therefore added a feature representing patient history and found that the performance of the *CombinedTagger* on *OSSD* increased 1.2% (results not reported in the Result section), although the increase was not statistically significant (*t* test, P=.25).

Limitations

Our study has limitations. The AETaggers were trained on the FAERS corpus we constructed. Like any other NLP system, the performance of the tagger on other types of EMRs can vary based on the structure and content of the narrative text. On the other hand, since our selection of the FAERS corpus was through a random process, we speculate that the data is representative. Although the taggers performed well, the training and evaluation was based on a relatively small training dataset. In the future, we would increase the size of the training data and explore other semi-supervised machine learning approaches to further improve the performance.

Conclusions

In this study, we developed an annotation guideline for medication and adverse event information from the FAERS narratives, our annotation of 122 FAERS narratives (a total of approximately 23,000 tokens) showed a reliable inter-rater annotation agreement (an overall kappa of .82). We then developed machine learning models for automatically extracting medication and adverse event information from the FAERS narratives. We explored utilizing different learning features in the machine learning models. The results showed that features such as syntactic, semantic, morphological, and affix improved the performance and the best performing system had an overall F1 score of 0.73. In the future, we would like to refine further the annotation guideline, explore additional features and increase the annotation size to improve system performance. We will also explore approaches for normalizing the entities by mapping them to standard terminologies like MedDRA and identify the causal relation between a medication and an adverse event.

Acknowledgments

We would like to acknowledge Dr Dennis W Raisch for providing the FAERS data from FDA and Ricky J Sethi for providing valuable comments on the manuscript. The research reported in this publication was supported in part by the National Institutes of Health (NIH) the National Institute of General Medical Sciences under award number 5R01GM095476 and the National Center for Advancing Translational Sciences under award number UL1TR000161. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

Conflicts of Interest

None declared.

Authors' Contributions

Conceived and designed the experiments: BPR HY. Performed Annotation: ZF NF SB. Performed the experiments: BPR. Obtained the data: DW. Analyzed the data: BPR HY. Wrote the paper: BPR HY.

Multimedia Appendix 1

Annotation Guideline.

[PDF File (Adobe PDF File), 183KB - medinform_v2i1e10_app1.pdf]

References

1. Kohn LT, Corrigan JM, Donaldson MS. To err is human: building a safer health system. In: To err is human: building a safer health system. Washington, D.C: National Academy Press; 2000.
2. E 2 A Clinical Safety Data Management: Definitions and Standards for Expedited Reporting - WC500002749. pdf URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002749.pdf [accessed 2013-10-01] [WebCite Cache ID 6K2qiistb]
3. Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, et al. Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ADE Prevention Study Group. JAMA 1995 Jul 5;274(1):29-34. [Medline: 7791255]
4. Nebeker JR, Barach P, Samore MH. Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting. Ann Intern Med 2004 May 18;140(10):795-801. [Medline: 15148066]
5. Forster AJ, Halil RB, Tierney MG. Pharmacist surveillance of adverse drug events. Am J Health Syst Pharm 2004 Jul 15;61(14):1466-1472. [Medline: 15332694]
6. Thomas EJ, Studdert DM, Burstin HR, Orav EJ, Zeena T, Williams EJ, et al. Incidence and types of adverse events and negligent care in Utah and Colorado. Med Care 2000 Mar;38(3):261-271. [Medline: 10718351]
7. Leape LL, Brennan TA, Laird N, Lawthers AG, Localio AR, Barnes BA, et al. The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study II. N Engl J Med 1991 Feb 7;324(6):377-384. [doi: 10.1056/NEJM199102073240605] [Medline: 1824793]
8. Classen DC, Pestotnik SL, Evans RS, Lloyd JF, Burke JP. Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality. JAMA 1997;277(4):301-306. [Medline: 9002492]
9. Cullen DJ, Sweitzer BJ, Bates DW, Burdick E, Edmondson A, Leape LL. Preventable adverse drug events in hospitalized patients: a comparative study of intensive care and general care units. Crit Care Med 1997 Aug;25(8):1289-1297. [Medline: 9267940]
10. Cullen DJ, Bates DW, Small SD, Cooper JB, Nemeskal AR, Leape LL. The incident reporting system does not detect adverse drug events: a problem for quality improvement. Jt Comm J Qual Improv 1995 Oct;21(10):541-548. [Medline: 8556111]

11. Bates DW, Spell N, Cullen DJ, Burdick E, Laird N, Petersen LA, et al. The costs of adverse drug events in hospitalized patients. Adverse Drug Events Prevention Study Group. JAMA 1997;277(4):307-311. [Medline: 9002493]
12. FDA Adverse Events Reporting System (FAERS) > FDA Adverse Event Reporting System (FAERS) Statistics. Available. Online URL: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070093.htm [accessed 2013-10-01] [WebCite Cache ID 6K2rK4ifG]
13. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med 2012 Mar 14;4(125):125ra31 [FREE Full text] [doi: 10.1126/scitranslmed.3003377] [Medline: 22422992]
14. Hiatt HH, Barnes BA, Brennan TA, Laird NM, Lawthers AG, Leape LL, et al. A study of medical injury and medical malpractice. N Engl J Med 1989 Aug 17;321(7):480-484. [doi: 10.1056/NEJM198908173210725] [Medline: 2761585]
15. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. N Engl J Med 1991 Feb 7;324(6):370-376. [doi: 10.1056/NEJM199102073240604] [Medline: 1987460]
16. Jha AK, Kuperman GJ, Teich JM, Leape L, Shea B, Rittenberg E, et al. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. J Am Med Inform Assoc 1998 Jun;5(3):305-314 [FREE Full text] [Medline: 9609500]
17. Murff HJ, Forster AJ, Peterson JF, Fiskio JM, Heiman HL, Bates DW. Electronically screening discharge summaries for adverse medical events. J Am Med Inform Assoc 2003 Aug;10(4):339-350 [FREE Full text] [doi: 10.1197/jamia.M1201] [Medline: 12668691]
18. Hwang SH, Lee S, Koo HK, Kim Y. Evaluation of a computer-based adverse-drug-event monitor. Am J Health Syst Pharm 2008 Dec 1;65(23):2265-2272. [doi: 10.2146/ajhp080122] [Medline: 19020194]
19. Tinoco A, Evans RS, Staes CJ, Lloyd JF, Rothschild JM, Haug PJ. Comparison of computerized surveillance and manual chart review for adverse events. J Am Med Inform Assoc 2011 Aug;18(4):491-497 [FREE Full text] [doi: 10.1136/amiajnl-2011-000187] [Medline: 21672911]
20. Gardner RM, Pryor TA, Warner HR. The HELP hospital information system: update 1998. Int J Med Inform 1999 Jun;54(3):169-182. [Medline: 10405877]
21. Hazlehurst B, Frost HR, Sittig DF, Stevens VJ. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. J Am Med Inform Assoc 2005 Oct;12(5):517-529 [FREE Full text] [doi: 10.1197/jamia.M1771] [Medline: 15905485]
22. Hazlehurst B, Naleway A, Mullooly J. Detecting possible vaccine adverse events in clinical notes of the electronic medical record. Vaccine 2009 Mar 23;27(14):2077-2083. [doi: 10.1016/j.vaccine.2009.01.105] [Medline: 19428833]
23. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc 2009 Jun;16(3):328-337 [FREE Full text] [doi: 10.1197/jamia.M3028] [Medline: 19261932]
24. Haerian K, Varn D, Vaidya S, Ena L, Chase HS, Friedman C. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. Clin Pharmacol Ther 2012 Aug;92(2):228-234 [FREE Full text] [doi: 10.1038/clpt.2012.54] [Medline: 22713699]
25. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc 2005 Aug;12(4):448-457 [FREE Full text] [doi: 10.1197/jamia.M1794] [Medline: 15802475]
26. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994 Apr;1(2):161-174 [FREE Full text] [Medline: 7719797]
27. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, et al. Extraction of adverse drug effects from clinical records. Stud Health Technol Inform 2010;160(Pt 1):739-743. [Medline: 20841784]
28. Yang C, Srinivasan P, Polgreen PM. Automatic adverse drug events detection using letters to the editor. AMIA Annu Symp Proc 2012;2012:1030-1039 [FREE Full text] [Medline: 23304379]
29. Gurulingappa H, Mateen-Rajput A, Toldo L. Extraction of potential adverse drug events from medical case reports. J Biomed Semantics 2012;3(1):15 [FREE Full text] [doi: 10.1186/2041-1480-3-15] [Medline: 23256479]
30. Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. AMIA Annu Symp Proc 2011;2011:217-226 [FREE Full text] [Medline: 22195073]
31. Nikfarjam A, Gonzalez GH. Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments. AMIA Annu Symp Proc 2011;2011:1019-1026 [FREE Full text] [Medline: 22195162]
32. White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. J Am Med Inform Assoc 2013 May 1;20(3):404-408 [FREE Full text] [doi: 10.1136/amiajnl-2012-001482] [Medline: 23467469]
33. Shetty KD, Dalal SR. Using information mining of the medical literature to improve drug safety. J Am Med Inform Assoc 2011 Oct;18(5):668-674 [FREE Full text] [doi: 10.1136/amiajnl-2011-000096] [Medline: 21546507]
34. Lindquist M. VigiBase, the WHO Global ICSR Database System: Basic Facts. Therapeutic Innovation & Regulatory Science September 2008 Sep;42(5):409-419. [doi: 10.1177/009286150804200501]
35. Hauben M, Madigan D, Gerrits CM, Walsh L, Van Puijenbroek EP. The role of data mining in pharmacovigilance. Expert Opin Drug Saf 2005 Sep;4(5):929-948. [doi: 10.1517/14740338.4.5.929] [Medline: 16111454]

36. Hauben M, Bate A. Decision support methods for the detection of adverse events in post-marketing data. Drug Discov Today 2009 Apr;14(7-8):343-357. [doi: 10.1016/j.drudis.2008.12.012] [Medline: 19187799]
37. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Ther 2012 Jun;91(6):1010-1021 [FREE Full text] [doi: 10.1038/clpt.2012.50] [Medline: 22549283]
38. Bate A, Evans SJ. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiol Drug Saf 2009 Jun;18(6):427-436. [doi: 10.1002/pds.1742] [Medline: 19358225]
39. Almenoff JS, Pattishall EN, Gibbs TG, DuMouchel W, Evans SJ, Yuen N. Novel statistical tools for monitoring the safety of marketed drugs. Clin Pharmacol Ther 2007 Aug;82(2):157-166. [doi: 10.1038/sj.clpt.6100258] [Medline: 17538548]
40. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001 Nov;10(6):483-486. [doi: 10.1002/pds.677] [Medline: 11828828]
41. van Puijenbroek EP, Bate A, Leufkens HG, Lindquist M, Orre R, Egberts AC. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf 2002 Feb;11(1):3-10. [doi: 10.1002/pds.668] [Medline: 11998548]
42. Norén GN, Sundberg R, Bate A, Edwards IR. A statistical methodology for drug-drug interaction surveillance. Stat Med 2008 Jul 20;27(16):3057-3070. [doi: 10.1002/sim.3247] [Medline: 18344185]
43. Harpaz R, Vilar S, Dumouchel W, Salmasian H, Haerian K, Shah NH, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. J Am Med Inform Assoc 2013 May 1;20(3):413-419 [FREE Full text] [doi: 10.1136/amiajnl-2012-000930] [Medline: 23118093]
44. Sakaeda T, Kadoyama K, Okuno Y. Adverse event profiles of platinum agents: data mining of the public version of the FDA adverse event reporting system, AERS, and reproducibility of clinical observations. Int J Med Sci 2011;8(6):487-491 [FREE Full text] [Medline: 21897761]
45. Vilar S, Harpaz R, Chase HS, Costanzi S, Rabadan R, Friedman C. Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis. J Am Med Inform Assoc 2011 Dec;18 Suppl 1:i73-i80 [FREE Full text] [doi: 10.1136/amiajnl-2011-000417] [Medline: 21946238]
46. Norén GN, Bate A, Orre R, Edwards IR. Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. Stat Med 2006 Nov 15;25(21):3740-3757. [doi: 10.1002/sim.2473] [Medline: 16381072]
47. Hochberg AM, Hauben M, Pearson RK, O'Hara DJ, Reisinger SJ, Goldsmith DI, et al. An evaluation of three signal-detection algorithms using a highly inclusive reference event database. Drug Saf 2009;32(6):509-525. [Medline: 19459718]
48. Bate A, Lindquist M, Edwards IR, Orre R. A data mining approach for signal detection and analysis. Drug Saf 2002;25(6):393-397. [Medline: 12071775]
49. Bate A. Bayesian confidence propagation neural network. Drug Saf 2007;30(7):623-625. [Medline: 17604417]
50. DuMouchel W, Pregibon D. Empirical bayes screening for multi-item associations. In: KDD-2001. New York, NY, USA: Association for Computing Machinery; 2001 Presented at: Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 26-29, 2001; San Francisco, CA, USA.
51. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol 1998 Jun;54(4):315-321. [Medline: 9696956]
52. Genkin A, Lewis DD, Madigan D. Large-Scale Bayesian Logistic Regression for Text Categorization. Technometrics 2007 Aug;49(3):291-304. [doi: 10.1198/004017007000000245]
53. Caster O, Norén GN, Madigan D, Bate A. Large-scale regression-based pattern discovery: The example of screening the WHO global drug safety database. Stat Anal Data Min 2010;3(4):197-208. [doi: 10.1002/sam.10078]
54. Solomon R, Dumouchel W. Contrast media and nephropathy: findings from systematic analysis and Food and Drug Administration reports of adverse effects. Invest Radiol 2006 Aug;41(8):651-660. [doi: 10.1097/01.rli.0000229742.54589.7b] [Medline: 16829749]
55. Rouane-Hacene M, Toussaint Y, Valtchev P. Mining Safety Signals in Spontaneous Reports Database Using Concept Analysis. In: AIME Proceedings / Lecture Notes in Artificial Intelligence. Berlin, Heidelberg: Springer; 2009 Presented at: 12th Conference on Artificial Intelligence in Medicine in Europe (AIME); July 18-22, 2009; Verona, Italy p. 285-294.
56. Harpaz R, Chase HS, Friedman C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. BMC Bioinformatics 2010;11 Suppl 9:S7 [FREE Full text] [doi: 10.1186/1471-2105-11-S9-S7] [Medline: 21044365]
57. Harpaz R, Perez H, Chase HS, Rabadan R, Hripcsak G, Friedman C. Biclustering of adverse drug events in the FDA's spontaneous reporting system. Clin Pharmacol Ther 2011 Feb;89(2):243-250 [FREE Full text] [doi: 10.1038/clpt.2010.285] [Medline: 21191383]
58. Ball R, Botsis T. Can network analysis improve pattern recognition among adverse events following immunization reported to VAERS? Clin Pharmacol Ther 2011 Aug;90(2):271-278. [doi: 10.1038/clpt.2011.119] [Medline: 21677640]
59. Harpaz R, DuMouchel W, LePendu P, Bauer-Mehren A, Ryan P, Shah NH. Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. Clin Pharmacol Ther 2013 Jun;93(6):539-546. [doi: 10.1038/clpt.2013.24] [Medline: 23571771]

60. Yom-Tov E, Gabrilovich E. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. J Med Internet Res 2013;15(6):e124 [FREE Full text] [doi: 10.2196/jmir.2614] [Medline: 23778053]
61. RADAR (Research on Adverse Drug events And Reports). Available. Online URL: http://www.feinberg.northwestern.edu/news/2013/07/radar.html [accessed 2013-10-01] [WebCite Cache ID 6K2rG1Z3L]
62. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010 Oct;17(5):514-518 [FREE Full text] [doi: 10.1136/jamia.2010.003947] [Medline: 20819854]
63. Li Z, Liu F, Antieau L, Cao Y, Yu H. Lancet: a high precision medication event extraction system for clinical text. J Am Med Inform Assoc 2010 Oct;17(5):563-567 [FREE Full text] [doi: 10.1136/jamia.2010.004077] [Medline: 20819865]
64. Knowtator. URL: http://knowtator.sourceforge.net/ [accessed 2014-05-06] [WebCite Cache ID 6PNDuSsKi]
65. Protégé. URL: http://protege.stanford.edu/ [accessed 2014-05-06] [WebCite Cache ID 6PNEHutVo]
66. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 1960;20(1). [doi: 10.1177/001316446002000104]
67. John L, Andrew M, Fernando C N P. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001 Presented at: 18th International Conference on Machine Learning (ICML); 2001; Williams College, Williamstown, MA, USA p. 282-289.
68. Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Available. Online URL: http://www.cs.waikato.ac.nz/ml/weka/ [accessed 2013-10-01] [WebCite Cache ID 6K2rPJb8r]
69. Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 2005 Jul 15;21(14):3191-3192 [FREE Full text] [doi: 10.1093/bioinformatics/bti475] [Medline: 15860559]
70. Simon T, Koller D. Support vector machine active learning with applications to text classification. The Journal of Machine Learning Research 2002;2:45-66. [doi: 10.1162/153244302760185243]
71. Thorsten J. Text categorization with support vector machines: Learning with many relevant features. In: Machine learning: ECML-98: Lecture Notes in Computer Science Volume 1398, 1998, pp 137-142. Berlin: Springer; 1998 Presented at: 10th European Conference on Machine Learning (ECML); April 1998; Chemnitz, Germany.
72. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput 2008:652-663 [FREE Full text] [Medline: 18229723]
73. Charniak E, Johnson M. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05 2005:173-180. [doi: 10.3115/1219840.1219862]
74. McClosky D. Ph. Brown University: . D. thesis, , , 2010 Any domain parsing: Automatic domain adaptation for natural language parsing URL: http://cs.brown.edu/research/pubs/theses/phd/2010/mcclosky.pdf [accessed 2013-10-01] [WebCite Cache ID 6K3Cssx17]
75. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21 [FREE Full text] [Medline: 11825149]
76. MetaMap - A Tool For Recognizing UMLS Concepts in Text. URL: http://metamap.nlm.nih.gov/ [accessed 2014-05-06] [WebCite Cache ID 6PNEdc6dD]
77. Agarwal S, Yu H. Biomedical negation scope detection with conditional random fields. J Am Med Inform Assoc 2010;17(6):696-701 [FREE Full text] [doi: 10.1136/jamia.2010.003228] [Medline: 20962133]
78. Agarwal S, Yu H. Detecting hedge cues and their scope in biomedical text with conditional random fields. J Biomed Inform 2010 Dec;43(6):953-961 [FREE Full text] [doi: 10.1016/j.jbi.2010.08.003] [Medline: 20709188]
79. Ramesh BP, Prasad R, Miller T, Harrington B, Yu H. Automatic discourse connective detection in biomedical text. J Am Med Inform Assoc 2012 Oct;19(5):800-808 [FREE Full text] [doi: 10.1136/amiajnl-2011-000775] [Medline: 22744958]
80. Marcus MP, Marcinkiewicz MA, Santorini B. Building a large annotated corpus of English: The Penn Treebank. Computational linguistics 1993;19(2).

Abbreviations

ABNER: A Biomedical Named Entity Recognizer
ADE: adverse drug event
AE: adverse event
CRF: Conditional Random Fields
EMR: Electronic Medical Records
FAERS: FDA Adverse Event Reporting System
FDA: Food and Drug Administration
IAA: inter-annotator agreement
MedDRA: Medical Dictionary for Regulatory Activities
ML: machine learning
NB: Naïve Bayes
NLP: natural language processing

OSSD: Other Signs, Symptoms and Diseases
POS: part of speech
SVM: Support Vector Machines

Edited by G Eysenbach; submitted 17.10.13; peer-reviewed by T Wadhwa, D Maclean, WC Su; comments to author 23.11.13; revised version received 10.12.13; accepted 10.12.13; published 27.06.14.

Please cite as:
Polepalli Ramesh B, Belknap SM, Li Z, Frid N, West DP, Yu H
Automatically Recognizing Medication and Adverse Event Information From Food and Drug Administration's Adverse Event Reporting System Narratives
JMIR Med Inform 2014;2(1):e10
URL: http://medinform.jmir.org/2014/1/e10/
doi:10.2196/medinform.3022
PMID:25600332

Viewpoint

Dynamic Clinical Data Mining: Search Engine-Based Decision Support

Leo Anthony Celi[1,2], MD, MSc, MPH; Andrew J Zimolzak[3], MD, MMSc; David J Stone[4], MD

[1]Harvard-MIT Division of Health Science and Technology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, United States

[2]Beth Israel Deaconess Medical Center, Boston, MA, United States

[3]Children's Hospital Informatics Program, Department of Pediatrics, Harvard Medical School, Boston, MA, United States

[4]University of Virginia School of Medicine, Departments of Anesthesiology and Neurosurgery, Charlottesville, VA, United States

Corresponding Author:
Leo Anthony Celi, MD, MSc, MPH
Harvard-MIT Division of Health Science and Technology
Institute for Medical Engineering and 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

Abstract

The research world is undergoing a transformation into one in which data, on massive levels, is freely shared. In the clinical world, the capture of data on a consistent basis has only recently begun. We propose an operational vision for a digitally based care system that incorporates data-based clinical decision making. The system would aggregate individual patient electronic medical data in the course of care; query a universal, de-identified clinical database using modified search engine technology in real time; identify prior cases of sufficient similarity as to be instructive to the case at hand; and populate the individual patient's electronic medical record with pertinent decision support material such as suggested interventions and prognosis, based on prior outcomes. Every individual's course, including subsequent outcomes, would then further populate the population database to create a feedback loop to benefit the care of future patients.

(JMIR Med Inform 2014;2(1):e13) doi:10.2196/medinform.3110

KEYWORDS

decision support; clinical informatics; big data

Introduction

With the near universal implementation of electronic medical records (EMRs) in conjunction with enhanced data storage options, the time nears for real time data utilization in the clinical care process [1]. The subject of the increasing importance of data for health care is much talked and written about, but there is much less discussion regarding how data might specifically be used to drive and improve the individual clinician-patient interactions that accrue to formulate the process of health care. In other words, how could complete clinical decision support be implemented across the entire health care system? Big data is an increasing presence in health care, but data of all sizes are still underutilized. In those instances when they are used at all, they are used mainly in a retrospective analytic manner to analyze outcomes, processes, and costs. Currently, they do not dynamically drive clinical decision making in real time.

We have written on the need for the better use of intensive care unit data, noting that the development of data-based clinical decision support (CDS) tools would be one of the benefits of more comprehensive data capture [2]. Currently, the medical digital world comprises systems that are technically networked, but with data that are not systematically gathered, captured, or analyzed [3]. There are several studies that have demonstrated the potential applications and potential of capturing and

analyzing clinical data [4,5]. In a more general response to this challenge, we describe a solution that combines the utilization of three fundamental components in real time: (1) big data, (2) search engines, and (3) EMRs. In particular, search engines are brilliant tools that we all utilize many times each day; however, they have not been systematically employed for the purpose of CDS, and they represent an overlooked resource.

Dynamic Clinical Data Mining

The struggle to implement EMRs is finally coming to a mainly successful end in North America. However, the current generation of EMRs serves to digitalize information, but not to leverage it. The next step in the clinical digitization process should be the creation of a medical Internet equivalent that incorporates the rapid, powerful data search engine features that all current Web users employ. We refer to the real time incorporation of external data into the workflow as dynamic clinical data mining (DCDM) (Figure 1 shows this mining). This process will drive the design of the next generation of EMRs, and it will subsequently support the next required stage of the digital transformation process by turning medical practice into a data driven, logical, and optimized system. This care support system will provide users with the timely information that they require to make the increasingly complex decisions of medical practice.

We propose a system in which the knowledge gained from the care of individuals systematically contributes to the care of populations. The loop is closed when the richer data available in the population datasets is subsequently used in the care of individuals. DCDM would leverage the automatic crowd sourcing available in the form of population outcome analysis to formulate individualized diagnostic and therapeutic recommendations in real time. In other words, our viewpoint aligns with the Committee on a Framework for Developing a New Taxonomy of Disease, who advocate that "researchers and health care providers have access to very large sets of health and disease-related data linked to individual patients" in order to facilitate precision medicine [6]. To the Committee's position, we would add our own that researchers and clinicians are already experienced with Internet search engines, so they would be comfortable with the identification of pertinent clinical information by accessing these large sets of data through a search engine metaphor. Currently, most clinical guidelines are generated by expert opinion based on experience and research findings such as randomized controlled trials [7]; DCDM would formulate the functional equivalent of personalized clinical guidelines.

While leading a team in the intensive care unit (ICU), one of us (LAC) experienced a difficult decision involving the resumption of anticoagulation in a patient with two mechanical heart valves. The patient was recovering from endocarditis complicated by brain abscesses. The team consulted local experts as well as the literature to guide them in weighing the risks and benefits of reinitiating anticoagulation, given the patient's age, comorbidities, the specific bacteria involved, the number of mechanical valves, the extent and current status of the infection, etc. The information resources that were accessed provided only general recommendations that were obviously not tailored to the patient's demographics and comorbidities, nor to the specifics of the clinical context. The majority of these recommendations were based on expert opinions or small clinical trials, and not on "gold-standard", multi-center randomized controlled studies. The decision was made to restart anticoagulation cautiously, given the patient's clinical stability, the absence of bleeding complications during the acute phase, and the lack of any planned surgical intervention. In fact, preparations were underway for discharge to a skilled nursing facility. Unfortunately, four days after reinitiation of anticoagulation, the patient suffered from a massive hemorrhage of one of the brain abscesses, prompting emergent hemicraniectomy. A DCDM system could have provided predictions of the harms and benefits of anticoagulation for such a complicated patient, and it would provide the previous outcomes associated with each treatment option to review in real time [8].

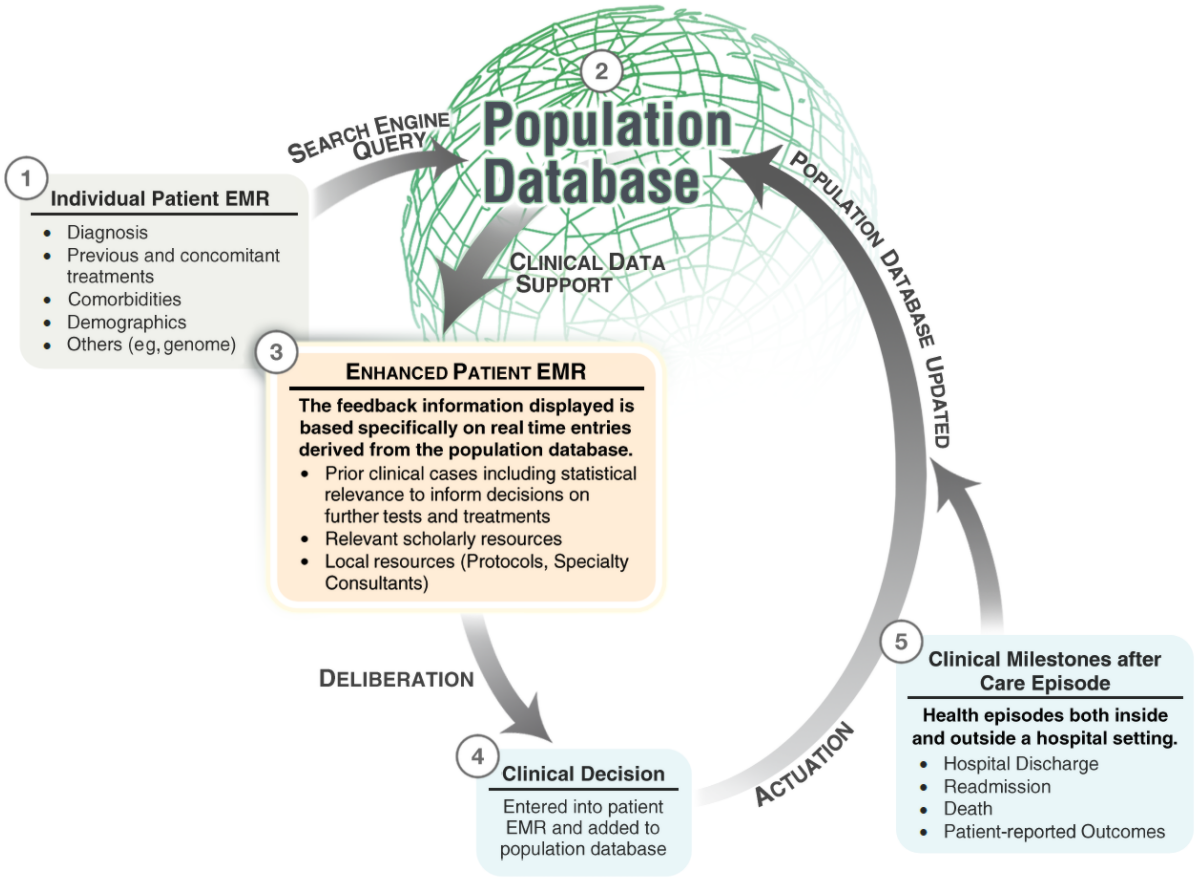
Uncertainties are not limited to complex scenarios, but occur with alarming frequency in all medical settings. For example, on a daily basis in the ICU, emergency department, or the operating room, clinicians target a desired blood pressure according to population-based guidelines. When hypotension ensues, the timing, mode, and extent of intervention to maintain that goal remain art rather than science. Given that interventions to raise blood pressure such as vasopressor therapy or fluids are associated with risk of harm if given even slightly in excess, it is crucial that the targeted blood pressure be personalized as much as possible. DCDM would add the knowledge gained from prior care of populations to the current local data specifics in order to formulate an approach that is optimal in terms of both the short-term goal as well as the long-term outcomes. For instance, DCDM could assist an ICU physician in choosing an intervention and its dose to treat shock, such that the intervention has the optimal effect on the short-term blood pressure profile and long-term mortality, length of stay, and/or eventual quality of life.

Other studies have explored similar themes. Certainly the application of logic and probability to medicine has been discussed for decades [9]. More recently, and more to the point of our discussion, a variety of commentators have called for a nationwide learning health system [10,11]. In 2011, Frankovich et al reported the case of a girl with lupus and potential thrombotic risk factors. To determine whether anticoagulation was appropriate, they used text searching to retrieve records of similar patients from their hospital's EMR, followed by a focused manual review. They found that pediatric patients with lupus and these potential risk factors indeed had a higher risk of thrombosis than those without the risk factors, and they elected to start anticoagulation as a result [12]. The Query Health initiative, from the Office of the National Coordinator for Health Information Technology, intends to facilitate distributed queries, which can aggregate results from multiple organizations' patient populations while preserving data security [13]. Similarly, the goal of the Strategic Health Information Technology Advanced Research Project on secondary use (SHARPn) is to standardize structured and unstructured EMR data to promote its reuse [14]. The open source Clinical Text Analysis and Knowledge

Extraction System and the SHARPn program use the Unstructured Information Management Architecture, the same architecture that allowed the Watson system from International Business Machines to compete on the Jeopardy! television quiz show [15,16]. In general, search engines for unstructured text are seen as the first step, and implementation of "content analytics" is the next step to extract information, allow exploration, and to improve search [17].

CDS provides caregivers with information to improve the quality of their decision making, yet caregivers still do not have available a dynamic, comparative analysis between the current patient and all available data generated during clinical care. This analysis is individually tailored because it uses EMR data entered on one specific patient, yet it remains population-based because the analysis makes a comparison to population data, to identify similar clinical situations from the past, and to mine them for interventions and subsequent outcomes (as illustrated in point 5 in Figure 1). Thus, the clinician does not have to make a decision in isolation from what has been tried, observed, and documented by many colleagues in many other similar patients. In addition, the information provided would provide useful support to the process of patient-physician shared decision making [18]. This approach would interrogate data to suggest next step options and weigh the risks and benefits of a treatment or test for a specific patient, the Holy Grail of personalized medicine.

Figure 1. Dynamic clinical data mining. Figure courtesy of Kai-ou Tang and Edward Moseley. EMR=electronic medical record.



Discussion

Required Data and Information

The most basic requirement for the DCDM system is the complete digital capture of patient information. We would maintain that de-identified clinical data constitutes a public good and should reside in a carefully managed public domain database, overseen by a cooperative coalition of vendors, provider institutions, and regulators. This is already the case for federally funded research data in the United States, and a movement is underway to share participant-level data from clinical trials [19-21]. Furthermore, the Patient Centered Outcomes Research Institute in the United States has already begun to develop the infrastructure that will aggregate large amounts of de-identified patient data from diverse sources for the purposes of observational research studies [22]. Any central database or federated query system must of course be governed by policies that account for the interests and preferences of the public regarding patient privacy, and the purposes for which the data are used [23]. The costs of database management would be built into purchase and maintenance agreements. Subsequent analyses would identify the clinical and financial impact of the entire data-based system with adjustments made as necessary.

In a DCDM system, a search engine would accept both structured and unstructured search terms to query the population database, much as current search engines query the database of

the entire Internet. The unstructured terms could be used in a query via real time natural language processing or the next generation of "text to code" conversion applications, which convert free text to coded, structured search terms, while considering the context provided by free text, in order to ensure accuracy and clinical intention. The individual's data would be rapidly compared to the population database to capture a set of useful records that match the content and context of the care encounter.

We envision every patient's health data digitally catalogued according to demographics, diagnoses, treatments, and outcomes, all time stamped for sequential interpretation. We suspect the types of data included will evolve rapidly over time. For example, future data may be derived from cell phones or home monitors. This will be the basis of a data-based learning system of care where choices are made on the basis of substantial data, statistical support programs, and documented outcomes, rather than on individual experience and inconsistent use of applicable informational resources.

Potential Obstacles

A significant caveat is that bias and/or residual confounding by indication may mar the analysis. The goal is to identify patient records in the database that are as similar as possible to the patient in terms of the variables that can confound the relationship between the intervention and the outcome as identified by clinician heuristics and complemented by computer algorithms, and then to compare the outcomes of those who receive the intervention versus those who did not. Residual confounding means that the outcome difference might not be due to the intervention, but rather due to something inherent to those patients who receive the treatment, or their condition. Realizing that the system is to be used by clinicians rather than data scientists, it must be designed so that such confounding and bias are minimized, with the confidence levels around the estimate of the treatment effects quantified and explained at the clinical user interface level.

The use of raw data from a variety of sources will present challenges. We acknowledge the inherent heterogeneity of people and disease. This presents an issue in terms of the levels of detail that require capture. The integration of data from multiple sources will require the use of standard terminologies and ontologies to allow for compatibility of the data from one source to another [14]. With the use of such standardization, these heterogeneities become inconveniences, not obstacles, to the vision. We foresee the implementation of progressively better EMRs, networks, and databases, all used by a generation of clinicians who have grown up with, are comfortable with, and expect to use and benefit from digital tools. It is important to anticipate potential risks, but this should be done in order to design and build the system so as to minimize them.

CDS tools must be engineered purposefully into workflow to avoid actually increasing user time and work requirements. An author of this paper (LAC) has previously reported on the use of local databases for the creation of CDS tools [24-26], which is one of the "grand challenges" in CDS [27]. Recent work adds the input of dynamic variables, which capture more information than traditional prediction models, including data on changes and variability of repeatedly measured values [28]. The readers are hereby directed to a recent review of the use of data mining in CDS [29]. DCDM would extend the capabilities of CDS by dynamically incorporating both individual and population data in real time [30,31]. In addition to querying and populating local databases, DCDM would also use the power of search engine technology to leverage population level data.

Organization and Actuation

Combined clinical and engineering teams would need to work together to generate algorithms to determine the weight of each feature being matched against the outcome of interest, as well as the relative value of (and permissible missing values for) the interacting data elements in the match process. These algorithms should be modifiable in order to meet the continuously changing practice of medicine. Search engine algorithms are modifiable and these modifications can be engineered for specific purposes. Google has made a number of such strategic modifications to its algorithms over time [32]. It is likely that a prototype employing a smaller search target such as the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) Database [33] would be required to demonstrate the practicality and utility of the concept, as well as to create, develop, and initially refine the search engine algorithm. Indeed, the MIMIC Database has been previously employed to predict fluid responsiveness among hypotensive patients [34], as well as the hematocrit trend among patients with gastrointestinal bleeding [35], using the trajectory of physiologic variables over time.

The system would identify and suggest prioritized interventions and other courses of action that have been shown to be most valuable in terms of outcome and cost. The system's features might include displays of quantitative and qualitative description of the match, hyperlinks that allow the user to drill further into the underlying data that is returned, and links to conventional practice guidelines and evidence-based modalities.

A clinical decision must be made at one point in time, but in most cases, decisions are ongoing and iterative. The system will incorporate the short-term response to the prior intervention each time that the system is accessed, but also capture long-term outcomes. For example, when a physician orders an intervention in response to acute kidney injury, the system would log the short-term response in serum creatinine, and also the long-term outcome of progression to or prevention of end-stage renal disease. The system could also be independently data mined to identify patterns that indicate whether the patient course is on track toward the desired outcome.

Large, diverse, international populations would improve the opportunity to achieve matches. When no match is possible, an alert could be provided noting the unusual features that preclude a match. The system would then provide appropriate suggestions for the user, such as a specialty referral, a data error of some kind, or even the possible detection of an entirely new condition. It would also serve as an epidemiological tool that recognizes emerging or spreading contagions [36,37], or other harmful exposures [38,39] more quickly and efficiently than is currently possible.

Conclusion

DCDM has its roots in the need for medical care to be more fully based on data. The universal collection of data would also present the additional advantages of providing future opportunities to formulate randomized registry trials, as well as for other directed data mining purposes [40]. DCDM would begin to transform the exigent data entries that clinicians perform on a daily basis into a real tool for clinical care. Decisions would be made on the basis of experience over vast populations, rather than solely on individual knowledge and experience. We propose the creation of a system that supports clinician decision makers so that their decisions can be as logical, transparent, and unambiguous as possible. DCDM would more gainfully employ the power of networked computers, search engines, and data storage advances to leverage the copious, but underused data entered into EMRs.

Acknowledgments

We would like to thank Marie Csete, MD, PhD; and Daniel Stone, PhD, for their helpful comments.

Conflicts of Interest

None declared.

References

1. Rodrigues JJ, de la Torre I, Fernández G, López-Coronado M. Analysis of the security and privacy requirements of cloud-based electronic health records systems. J Med Internet Res 2013;15(8):e186 [FREE Full text] [doi: 10.2196/jmir.2494] [Medline: 23965254]
2. 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 [FREE Full text] [doi: 10.1164/rccm.201212-2311ED] [Medline: 23725609]
3. The Committee on the Learning Health Care System in America. In: Smith M, Saunders R, Stuckhardt L, McGinnis JM, editors. Best care at lower cost: The path to continuously learning health care in America. USA: The National Academies Press; 2013.
4. Fine AM, Nigrovic LE, Reis BY, Cook EF, Mandl KD. Linking surveillance to action: Incorporation of real-time regional data into a medical decision rule. J Am Med Inform Assoc 2007;14(2):206-211 [FREE Full text] [doi: 10.1197/jamia.M2253] [Medline: 17213492]
5. Davis DA, Chawla NV, Christakis NA, Barabási A. Time to CARE: A collaborative engine for practical disease prediction. Data Min Knowl Disc 2009 Nov 25;20(3):388-415. [doi: 10.1007/s10618-009-0156-z]
6. Committee on a Framework for Development a New Taxonomy of Disease, Board on Life Sciences, Division on Earth and Life Studies, National Research Council. Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease. USA: National Academies Press; 2011.
7. Carville S, Harker M, Henderson R, Gray H, Guideline Development Group. Acute management of myocardial infarction with ST-segment elevation: Summary of NICE guidance. BMJ 2013;347:f4006. [Medline: 23843545]
8. Bennett C, Doub T. caseybennett.com. 2010. Data mining and electronic health records: Selecting optimal clinical treatments in practice URL: http://www.caseybennett.com/uploads/DMIN_Paper_Proposal_Final__Arxiv.pdf [accessed 2014-06-14] [WebCite Cache ID 6QKOCNatb]
9. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science 1959 Jul 3;130(3366):9-21. [Medline: 13668531]
10. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med 2010 Nov 10;2(57):57cm29. [doi: 10.1126/scitranslmed.3001456] [Medline: 21068440]
11. Roundtable on Value & Science-Driven Health Care, The Learning Health System Series, Institute of Medicine. In: Grossman C, McGinnis JM, editors. Digital infrastructure for the learning health system: The foundation for continuous improvement in health and health care: Workshop series summary (the learning health system series). USA: National Academies Press; 2011.
12. Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. N Engl J Med 2011 Nov 10;365(19):1758-1759. [doi: 10.1056/NEJMp1108726] [Medline: 22047518]
13. Klann JG, Buck MD, Brown J, Hadley M, Elmore R, Weber GM, et al. Query health: Standards-based, cross-platform population health surveillance. J Am Med Inform Assoc 2014 Jul;21(4):650-656. [doi: 10.1136/amiajnl-2014-002707] [Medline: 24699371]
14. Pathak J, Bailey KR, Beebe CE, Bethard S, Carrell DC, Chen PJ, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: The SHARPn consortium. J Am Med Inform Assoc 2013 Dec;20(e2):e341-e348. [doi: 10.1136/amiajnl-2013-001939] [Medline: 24190931]
15. Chute CG, Pathak J, Savova GK, Bailey KR, Schor MI, Hart LA, et al. The SHARPn project on secondary use of electronic medical record data: Progress, plans, and possibilities. AMIA Annu Symp Proc 2011;2011:248-256 [FREE Full text] [Medline: 22195076]

16. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
17. Feldman S, Reynolds H, Schubmehl D. Content analytics and the high-performing enterprise. Framingham, MA: IDC; 2012. URL: http://public.dhe.ibm.com/software/data/sw-library/ecm-programs/ContentAnalytics_IDC.pdf [accessed 2014-06-05] [WebCite Cache ID 6Q6xdxLqK]
18. Oshima Lee E, Emanuel EJ. Shared decision making to improve care and reduce costs. N Engl J Med 2013 Jan 3;368(1):6-8. [doi: 10.1056/NEJMp1209500] [Medline: 23281971]
19. Holdren JP. Executive Office of the President, Office of Science and Technology Policy. Increasing access to the results of federally funded scientific research Internet URL: http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf [accessed 2013-11-18] [WebCite Cache ID 6LEYusHy6]
20. Mello MM, Francer JK, Wilenzick M, Teden P, Bierer BE, Barnes M. Preparing for responsible sharing of clinical trial data. N Engl J Med 2013 Oct 24;369(17):1651-1658. [doi: 10.1056/NEJMhle1309073] [Medline: 24144394]
21. Drazen JM. Open data. N Engl J Med 2014 Feb 13;370(7):662. [doi: 10.1056/NEJMe1400850] [Medline: 24450856]
22. Selby J. Patient-centered outcomes research institute. Building PCORnet as a secure platform for outcomes research URL: http://www.pcori.org/blog/building-pcornet-as-a-secure-platform-for-outcomes-research/ [accessed 2014-05-13] [WebCite Cache ID 6PXClkL4I]
23. 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]
24. Lee J, Kothari R, Ladapo JA, Scott DJ, Celi LA. Interrogating a clinical database to study treatment of hypotension in the critically ill. BMJ Open 2012;2(3) [FREE Full text] [doi: 10.1136/bmjopen-2012-000916] [Medline: 22685222]
25. Celi LA, Tang RJ, Villarroel MC, Davidzon GA, Lester WT, Chueh HC. A clinical database-driven approach to decision support: Predicting mortality among patients with acute kidney injury. J Healthc Eng 2011 Mar;2(1):97-110 [FREE Full text] [doi: 10.1260/2040-2295.2.1.97] [Medline: 22844575]
26. Celi LA, Galvin S, Davidzon G, Lee J, Scott D, Mark R. A database-driven decision support system: Customized mortality prediction. J Pers Med 2012 Sep 27;2(4):138-148 [FREE Full text] [doi: 10.3390/jpm2040138] [Medline: 23766893]
27. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, et al. Grand challenges in clinical decision support. J Biomed Inform 2008 Apr;41(2):387-392 [FREE Full text] [doi: 10.1016/j.jbi.2007.09.003] [Medline: 18029232]
28. Mayaud L, Lai PS, Clifford GD, Tarassenko L, Celi LA, Annane D. Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension. Crit Care Med 2013 Apr;41(4):954-962 [FREE Full text] [doi: 10.1097/CCM.0b013e3182772adb] [Medline: 23385106]
29. Aleksovska-Stojkovska L, Loskovska S. Data mining in clinical decision support systems. In: Gaol FL, editor. Recent progress in data engineering and internet technology. Berlin: Springer; 2013:287-293.
30. Friedman CP, Rubin JC. The learning system: Now more than ever - Let's achieve it together!. 2014 Feb 25 Presented at: HIMSS 14 Annual Conference & Exhibition; 2014-02-25; Orlando, Florida URL: http://himss.files.cms-plus.com/2014Conference/handouts/70.pdf
31. Clinical personalized pragmatic predictions of outcomes. Department of Veterans Affairs URL: http://www.clinical3po.org/ [accessed 2014-05-13] [WebCite Cache ID 6PXBjQHPN]
32. Sullivan D. Search Engine Land. FAQ: All about the new Google "hummingbird" algorithm URL: http://searchengineland.com/google-hummingbird-172816 [accessed 2013-11-18] [WebCite Cache ID 6LEYRMJjZ]
33. PhysioNet. 2012. MIMIC II Databases URL: http://physionet.org/mimic2/ [accessed 2013-11-18] [WebCite Cache ID 6LEYagQz1]
34. Fialho AS, Celi LA, Cismondi F, Vieira SM, Reti SR, Sousa JM, et al. Disease-based modeling to predict fluid response in intensive care units. Methods Inf Med 2013;52(6):494-502. [doi: 10.3414/ME12-01-0093] [Medline: 23986268]
35. Cismondi F, Celi LA, Fialho AS, Vieira SM, Reti SR, Sousa JM, et al. Reducing unnecessary lab testing in the ICU with artificial intelligence. Int J Med Inform 2013 May;82(5):345-358. [doi: 10.1016/j.ijmedinf.2012.11.017] [Medline: 23273628]
36. Gesteland PH, Gardner RM, Tsui FC, Espino JU, Rolfs RT, James BC, et al. Automated syndromic surveillance for the 2002 Winter Olympics. J Am Med Inform Assoc 2003;10(6):547-554 [FREE Full text] [doi: 10.1197/jamia.M1352] [Medline: 12925547]
37. Elkin PL, Froehling DA, Wahner-Roedler DL, Brown SH, Bailey KR. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. Ann Intern Med 2012 Jan 3;156(1 Pt 1):11-18. [doi: 10.7326/0003-4819-156-1-201201030-00003] [Medline: 22213490]
38. Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse events using information technology. J Am Med Inform Assoc 2003;10(2):115-128 [FREE Full text] [Medline: 12595401]
39. 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]
40. Lauer MS, D'Agostino RB. The randomized registry trial--the next disruptive technology in clinical research? N Engl J Med 2013 Oct 24;369(17):1579-1581. [doi: 10.1056/NEJMp1310102] [Medline: 23991657]

Abbreviations

CDS: clinical decision support
DCDM: dynamic clinical data mining
EMRs: electronic medical records
ICU: intensive care unit
MIMIC: Multiparameter Intelligent Monitoring in Intensive Care
SHARPn: Strategic Health Information Technology Advanced Research Project on secondary use

Edited by G Eysenbach; submitted 19.11.13; peer-reviewed by D Maslove, S Seevanayanagam, J Klann, A Holzinger, WC Su, J Turner; comments to author 27.01.14; revised version received 28.02.14; accepted 22.05.14; published 23.06.14.

Please cite as:
Celi LA, Zimolzak AJ, Stone DJ
Dynamic Clinical Data Mining: Search Engine-Based Decision Support
JMIR Med Inform 2014;2(1):e13
URL: http://medinform.jmir.org/2014/1/e13/
doi:10.2196/medinform.3110
PMID:25600664

Review

Big Data and Clinicians: A Review on the State of the Science

Weiqi Wang[1], PhD; Eswar Krishnan[1], MD, MPH

School of Medicine, Stanford University, Palo Alto, CA, United States

Corresponding Author:
Eswar Krishnan, MD, MPH
School of Medicine
Stanford University
1000 Welch Road, Suite 203
Palo Alto, CA, 94304
United States
Phone: 1 650 725 8004
Fax: 1 650 723 9656
Email: e.krishnan@stanford.edu

Abstract

Background: In the past few decades, medically related data collection saw a huge increase, referred to as big data. These huge datasets bring challenges in storage, processing, and analysis. In clinical medicine, big data is expected to play an important role in identifying causality of patient symptoms, in predicting hazards of disease incidence or reoccurrence, and in improving primary-care quality.

Objective: The objective of this review was to provide an overview of the features of clinical big data, describe a few commonly employed computational algorithms, statistical methods, and software toolkits for data manipulation and analysis, and discuss the challenges and limitations in this realm.

Methods: We conducted a literature review to identify studies on big data in medicine, especially clinical medicine. We used different combinations of keywords to search PubMed, Science Direct, Web of Knowledge, and Google Scholar for literature of interest from the past 10 years.

Results: This paper reviewed studies that analyzed clinical big data and discussed issues related to storage and analysis of this type of data.

Conclusions: Big data is becoming a common feature of biological and clinical studies. Researchers who use clinical big data face multiple challenges, and the data itself has limitations. It is imperative that methodologies for data analysis keep pace with our ability to collect and store data.

(JMIR Med Inform 2014;2(1):e1) doi:10.2196/medinform.2913

KEYWORDS

big data; database; medical informatics; clinical research; medicine

Introduction

Big data refers to very large datasets with complex structures that are difficult to process using traditional methods and tools. The term process includes, capture, storage, formatting, extraction, curation, integration, analysis, and visualization [1-9]. A popular definition of big data is the "3V" model proposed by Gartner [10], which attributes three fundamental features to big data: high volume of data mass, high velocity of data flow, and high variety of data types. The notion of big data can be traced back to the 1970s [11-13] when scientists realized that they lacked the tools to analyze datasets of large size. In those days, big data was merely several to hundreds of megabytes [14]; now datasets of terabytes are common [15, 16]. Therefore, the "big" in big data reflects the limits of data storage and computational power existing at a given point in time.

Table 1 shows the growth of global big data volume and computer science papers on big data since 2009. This table exemplifies that stored data will be in the tens of zettabytes range by 2020, and research on how to deal with big data will grow exponentially as well.

Big data is gathered in many disciplines and is made possible by ubiquitous information-sensing devices and software [19]. An example is web logs: websites such as Google or Facebook automatically record user information at each visit. Other examples come from the stock market [20], earthquake

surveillance [21], political elections [22], behavioral studies [23], sports [24], pharmaceutical reports [25], health care [26, 27], electronic medical records [28], imaging data [29], genome data [30, 31], and entrepreneur transaction records [32]. Data collection is sometimes interdisciplinary. As an example, a sudden increase in Google search terms such as "flu symptoms" and "flu treatments" can be used to predict an increase in flu patients presenting to hospital emergency rooms [33]. This example also demonstrates that big data has promising predictive power and return on investment. Return on investment of big data has also been suggested for clinical big data [34, 35].

Although arguably valuable, big data is difficult to analyze due to the massive volume of the raw data and its diversity, as shown in Figure 1. Therefore, instead of the raw big data, a large dataset is often extracted from the raw data to generate a secondary storage of data for analysis purposes. This data extraction is applied, for example, when a computer tomography scan is involved in clinical trials and only the physician diagnosis based on the scan is included in data analysis. Similarly, a large volume of descriptive data on various kinds of samplings, tests, or assays can be extracted with only the key parameters kept. As a consequence, the data analyzed in clinical medicine is usually from secondary datasets that contain only data of interest. The secondary datasets, although still large, are not terabytes in size. Additionally, due to the nature of clinical trials, a large dataset in clinical medicine usually does not have an overwhelming number of samples. Kjaergard et al [36] defined clinical trials with 1000 or more participants as large, and the studies in clinical medicine titled big/large, data/dataset generally have thousands of attributes, but only hundreds of samples [37-39].

For this paper, we reviewed the literature to determine the features of clinical big data and determine the methods used for manipulation and analysis of these data. This paper is focused on clinical medicine rather than general health care issues; therefore, we mainly reviewed the studies that appeared relevant to clinicians. We examined the selected studies to extract information on research interests, goals, and achievements, and the implemented methodologies. Our intention was not to conduct an exhaustive systematic review, but instead to enable a literature-based discussion of how the big data issue has been addressed in clinical medicine. Based on our findings, we discuss the challenges and limitations of analysis of large clinical datasets.

Table 1. Global growth of big data and computer science papers on big data.

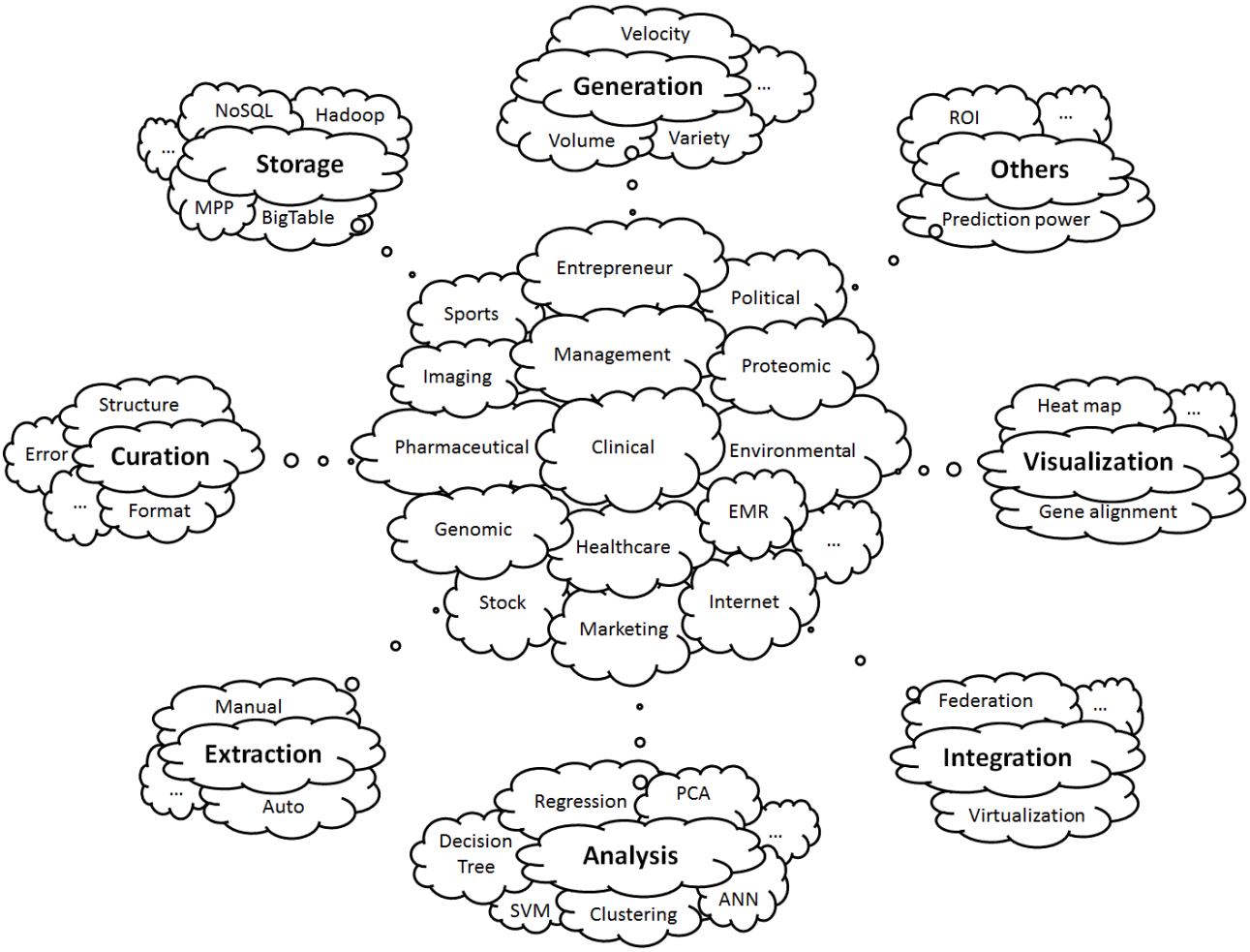
Year	Data volume, ZB[a,c]	Conference papers, CS[b,c]	Journal papers, CS[c]
2009	1.5	12	7
2010	2	26	7
2011	2.5	32	23
2012	3	78	47
2015	8	?	?
2020	44	??	??

[a]Data from *oracle* [17].

[b]Data from *Research Trends* [18].

[c]CS, computer science; ZB, zettabytes (1 zettabyte = 1000 terabytes = 10^6 petabytes = 10^{18} gigabytes, GB).

Figure 1. A schematic of the issues surrounding storage and use of big data. Clinical big data, as well as big data in other disciplines, have been surrounded by a number of issues and challenges, including (but not limited to): generation, storage, curation, extraction, integration, analysis, visualization, etc. ANN: artificial neuron network; EMR: electronic medical record; MPP: massively parallel-processing; PCA: principle component analysis; ROI: return of investment; SVM: support vector machine.



Methods

We conducted a literature review to identify studies on big data in medicine, especially clinical medicine. We used different combinations of keywords to search PubMed, Science Direct, Web of Knowledge, and Google Scholar for literature of interest, mainly from the last 10 years. The key words were: "big data medicine", "large dataset medicine", “clinical big data”, “clinical large dataset”, “clinical data warehouse”, “clinical database”, “clinical data mining”, “biomedical big data”, “biomedical database”, “biomedical data warehouse”, “healthcare big data”, “healthcare database”, and “healthcare data warehouse”.

Results

Big Data in Clinical Medicine

Big data plays an important role in medical and clinical research and has been leveraged in clinically relevant studies. Major research institute centers and funding agencies have made large investments in the arena. For example, the National Institutes of Health recently committed US $100 million for the big data to Knowledge (BD2K) initiative [40]. The BD2K defines “biomedical” big data as large datasets generated by research groups or individual investigators and as large datasets generated by aggregation of smaller datasets. The most well-known examples of medical big data are databases maintained by the Medicare and Healthcare Cost and Utilization Project (with over 100 million observations). One of the differences between medical big data and large datasets from other disciplines is that clinical big data are often collected based on protocols (ie, fixed forms) and therefore are relatively structured, partially due to the extraction process that simplify raw data as mentioned above. This feature can be traced back to the Framingham Heart Study [41], which has followed a cohort in the town of Framingham, Massachusetts since 1948. Vast amounts of data have been collected through the Framingham Heart Study, and the analysis has informed our understanding of heart diseases, including the effects of diet, exercise, medications, and obesity on risk [42]. There are many other clinical databases with different scopes, including but not limited to, prevalence and trend studies, risk factor studies, and genotype-phenotype studies.

Prevalence and Trend Studies

One of the major uses for clinical big data is in analysis of the prevalence or trends of a disease or phenotype among different populations. An early big data study evaluated a cohort consisting of 890,394 US veterans with diabetes followed from 2002 through 2006 [43]. Bermejo-Sanchez et al [44] observed

326 of the birth defect Amelia among 23 million live births, stillbirths, and fetal anomalies from 23 countries and 4 continents, and found the trend of higher prevalence of Amelia among younger mothers. Histological features that differ between chronic idiopathic inflammatory bowel disease and normality and between Crohn's disease and ulcerative colitis were identified in 809 large bowel endoscopic biopsies [45]. Kelly et al [46] studied the prevalence of hip abnormalities of 8192 subjects with hemophilia A and B. Siregar et al [47] performed a population-based study on patients after cardiac surgery in all 16 cardiothoracic surgery centers in the Netherlands. Elshazly et al [48] examined 1.3 million US adults for patient-level discordance of non-high-density lipoprotein cholesterol and low-density lipoprotein cholesterol. Chan and McGarey [49] summarize how large datasets can be analyzed to achieve population-based conclusions, specifically for determination of secular trends, health disparities, geographic variation, and evaluation of specific diseases and treatments. This paper also summarized the strengths and limitations of large-sized datasets and addressed issues such as missing data and bias. These issues will also be discussed in brief below.

Risk Factor Studies

Clinical big data can also be used to determine causality, effect, or association between risk factors and the disease of interest. Ursum et al [50] examined the relationships between seroconversion and patient age with inflammatory effects of autoantibodies in 18,658 rheumatoid arthritis patients and controls, and showed that citrullinated proteins and peptides were more reliable markers for rheumatoid arthritis than was Immunoglobulin M rheumatoid factor. Ajdacic-Gross et al [51] examined the data on 11,905 Swiss conscripts from 2003 for stuttering and found that there was no single overwhelming risk factor for stuttering, although premature birth and parental alcohol abuse appeared influential. Data collected on 14,433 patients from the 155 Veterans Administration medical centers in all 50 US states, Puerto Rico, and the District of Columbia were used to identify the alcohol dependence of medications [52]. By analysis of 53,177 cases of contrast administration in 35,922 patients from the Radiology and Cardiac Catheterization Laboratory databases, an increase in contrast nephropathy was associated with use of sodium bicarbonate [53]. Echocardiography and electrocardiogram-gated single-photon emission computed tomography traces for the evaluation of left ventricular ejection fraction were compared in 534 patients [54]. Zhang et al [55] examined clinical data of 16,135 adult patients and elucidated the relationships between glycemic, blood glucose level, and intake of insulin with mortality. Mitchel et al [56] studied the effect of 2 types of insulin on 7720 patients selected from 8 million in UK. Kobayashi et al [57] analyzed 19,070 records on right hemicolectomy from 3500 Japanese hospitals and successfully developed a risk model. It should be noted that in these studies, the terms of "association" and "causality" must be rigorously distinguished; most of the studies claimed association, whereas causality was rarely asserted.

Genotype–Phenotype Studies

With the advancement of genotyping technology, an increasing amount of risk-factor studies have attempted to assess association on the genetic level through evaluation of gene expression and/or genomic data obtained from patients and controls. For example, clinical and genetic data from 5700 patients who had been treated with warfarin were used to create an algorithm to estimate the appropriate dose [58]. Causality of autism spectrum disorders has been investigated by analysis of 31,516 clinical cases on copy number variation in patients versus 13,696 controls [59]. Koefoed et al [60] made efforts to assess the effects of signal transmission and calculated all combinations of three genotypes from 803 single-nucleotide polymorphism (SNP) genotypes (2.3 billion combinations) for 1355 controls and 607 patients with bipolar disorder. These studies are similar to risk-factor studies, yet often the big data is significantly larger in volume due in genetic analyses than in risk-factor studies.

Method Development Studies

A number of studies have taken advantage of clinical big data to establish new methods or techniques, or to develop new tools to enable analysis of data and decision making. In a typical example, Hill et al [61] designed an interface to use clinical data to evaluate risk ratios for various diseases to aid in evaluation of treatment options. Liu et al [62, 63] have used large-scale data analysis to optimize diagnosis of breast cancer from full-field digital mammography images. Lin et al [64] made efforts to formalize the phenotype variable in the database Genotypes and Phenotypes. Stephen et al [65] developed an algorithm to categorize pediatric patients presenting with respiratory distress into different subtypes using clinical variables from a clinical data warehouse. Clinical data warehouses or databases have been created from radiotherapy clinical trial data [66], gene mutations [67], cancer patient data [68, 69], kidney disease patient data [70], and gastrointestinal surgery patient data [71]. Additionally, studies have focused on personalized big data [72], citizen-centric health care versus patient-centric health care [72, 73], medication orders [74, 75], and decision making and information management/retrieval in general [75-80]. The dramatic increase in the number of studies with large scope in the past few years indicates an increasing desire of researchers to manipulate clinical big data; "big data-assisted clinics" may be expected in the near future.

Discussion

Diversity of Data in Clinical Medicine

The huge body of medical research that has been performed using large datasets demonstrates the broad spectrum of data resources used and shows that the structure of the medical dataset depends on the research question. Data from different subareas of medical research have broad diversity in terms of numbers of entries, types of data stored (or levels), dimensionality, and sample size [81]. Datasets obviously differ greatly in size: gene expression datasets derived from high-throughput microarray and next-generation sequencing technologies, such as those that analyze SNPs and copy number variations, tend to be massive, whereas clinical trial dataset are not as big. Phan et al [82] suggested that data in medicine be divided into four different levels: the molecular level (eg, genome data), cellular and tissue level (eg, stem cell

differentiation data), clinical and patient level (eg, clinical trial data), and biomedical knowledge base level (ie, a comprehensive data collection). Additionally, data tend to have different levels of dimensionality (ie, number of attributes or parameters, *p*) and sample sizes (ie, number of records/entries, n). Typical datasets fall into one of three categories, as summarized by Sinha et al [83]: large n, small *p*; small n, large *p*; and large n, large *p*. Thanks to advancements in computational technology, most algorithms can handle low-dimensional data (ie, large n, small *p*) without encountering significant difficulty.

Most clinical data, however, is high-dimensional (ie, small n, large *p* or large n, large *p*) due to a limited number of patients. One typical example comes from a study of 69 Broca's aphasic patients (ie, n=69) who were tested with nearly 6000 stimulus sentences (ie, *p*~6000) [84]. With similar dimensionality, Mitchell et al [39] studied bipolar disorder where the sample consisted of only 217 patients. For high-dimensional data, each point, sample, or element is described by many attributes [83] with the involvement of the "curse of dimensionality" [85]. Because high-dimensional data are sparse in dimensions, most classification or clustering approaches do not work well because the increase in problem space reduces the overall density of data samples. To solve this problem, compression methods and significance testing are usually used to either reduce the dimensionality or select relevant features before data analysis by some sort of data preprocessing [83].

Methods for Manipulation of Clinical Big Data

Technologies for Data Storage and Handling

Due to the massiveness and complexity of big data, nonrelational and distributed databases such as Apache Hadoop [86], Google BigTable [87], NoSQL [88], and massively parallel-processing databases are used rather than traditional relational databases to store data. A large number of biostatistics software packages have been used to handle large clinical datasets, some of which enabled the features of cloud-based or distributed computing. Popular software packages include, but are not limited to, SAS [36, 51-53], Mplus [51], SPSS [36, 39, 45], PP-VLAM [89], Stata [90], and R [91]. These technologies and tools greatly facilitate the handling of big data.

Methodologies for Data Preprocessing

Clinical raw big data can be highly diverse and uninformative without preprocessing. Extraction of a diagnosis from raw computer tomography data is an example of one of the predominant manners in which clinical big data are preprocessed. This type of processes relies on a specialist's personal expertise and can be a source of bias. Most early analyses of big data, including that collected by the Framingham Heart Study adopted some form of preprocessing; therefore, challenges exist in curation [6]. As an alternative to expert preprocessing, computational algorithms or statistical approaches, including compression methods, significance testing, or normalization [92] can be implemented to preprocess raw big data. This methodology may also introduce bias and can cause uncertainty problems during data integration.

In some scenarios, visualization can be a part of data preprocessing (as well as result exhibition). Typical examples in this regard include the use of heat maps [93], gene alignments [94], protein structure visualization [95], scatterplot matrix, tree visualization, network visualization, parallel coordinates, stacked graphs, etc. When the big data of interest are scattered or stored at different resources, data integration [96, 97] and federation [98] is an important phase during data preprocessing. Approaches such as the Information Manifold [97], which allows browsing and querying of multiple networked information sources, can provide solutions to uncertainty problems after data integration and mapping [99].

Statistical Approaches to Data Analysis

A number of popular statistical methods have been implemented in clinical data analysis. The most common include linear regression and logistic regression [30], latent class analysis [100], principle component analysis [101], and classification and regression trees [100]. Additionally, logarithmic and square-root transformations [58], naive Bayes methods [102], decision trees [103], neural networks [104], support vector machines [105], and hidden Markov models [83] are also used to study problems in medical data.

When a dataset is not overly complicated, a single test (eg, a simple Student's *t* test) should be powerful enough to reject a null hypothesis, and single hypothesis testing is the methodology to adopt [106]. Sometimes one cannot establish the significance of a hypothesis until different statistical tests have been applied to the same dataset. Multiple testing is often used to identify correlations that deserve further investigation [107]. Algorithms for false discovery rate [108] and family-wise error rate [109] calculation have been implemented for multiple testing in studies on gene expression data and datasets with similar levels of complexity.

Challenges and Limitations of Use of Clinical Big Data

Overview

Big data itself has many limitations. These limitations include "adequacy, accuracy, completeness, nature of the reporting sources, and other measures of the quality of the data", as summarized previously [110]. The consequences of these limitations are succinctly summarized in the book titled "Models. Behaving. Badly." [111]. Modeling can often lead to a biased statistical correlation or inference, sometimes known as a "false discovery". Clinical big data users face a large spectrum of challenges, including but not limited to sample size, selection bias, interpretation problem, missing values, dependence problems, and data handling methodologies.

Sample Size

One of the counterintuitive challenges in analysis of big data clinical datasets is that sometimes the sample size is not as big, compared with the number of attributes to allow statistically significant analysis. Population survey methods are sometimes adopted because these methods can provide larger datasets. However, the authenticity and accuracy of this type of data are arguably limited; hence, survey methods cannot be reliably used to produce an adequate description or prediction [39].

Selection Bias

Any dataset is a selection of data rather than the whole data world; therefore, selection bias is a very real limitation [112] even if the sample size is big. In that sense, all studies of clinical data have this limitation to some degree [39].

Interpretation Problem

Gebregziabher et al [43] stated that the datasets generated through many translational research projects to answer questions of public health interest are not self-explanatory due to complexity and inadequate description/documentation of the dataset's parameters and associated metadata. The methodologies for interpreting the data can therefore be subject to all sorts of philosophical debate. For example, the data may not be totally naïve or objective and interpretation may be biased by subjective assumptions and/or manipulations by individual analysts.

Missing Values

It is common problem that large datasets have missing values, and in many cases the problem can be significant [44]. A typical example is the Framingham Heart Study where data on serum uric acid are largely missing. Additionally, the covariates (ie, attributes) may not fully capture the degree of risk for patients and may cause uncertainty in results [53].

Dependence Problems

One issue that has been often neglected is the dependence of data. Dependence between either attributes or samples in datasets can cause the degrees of freedom to decrease and/or some statistical principles to no longer apply. Examples of this are found when the same patients are evaluated multiple times through follow-up and when correlations in gene expression are drawn based on samples from different patients treated with similar medications [83]. As many statistical methods do not account for dependence, results from these tests may be unreliable if this issue is not properly addressed before the data analysis.

Data Handling Methodologies

Effective processing of big data has always been a challenge. One must consider all the aspects of the dataset, including collection, curation, extraction, integration, interpretation, imputation, and selection of appropriate statistical methods, during processing and analysis. It has been claimed that analyses of large datasets are often suboptimal due to the researcher's lack of knowledge of the available tools and methodologies [83]. On the other hand, algorithms to handle big data are also underdeveloped to some extent and deserve more attention [113].

Conclusions

This paper reviewed studies that analyzed clinical big data and that discuss issues related to data storage and analysis. Big data is becoming a common feature of biological and clinical studies. Today, a single biophysical researcher can generate terabytes of data in hours. Over the last decade, clinical datasets have grown incredibly in size, mostly due to use of modern technologies for collection and recording of data. Researchers who use clinical big data face multiple challenges, and the data itself has limitations. It is imperative that methodologies for data analysis keep pace with our ability to collect and store data.

Authors' Contributions

WW did the search and primary review of the literature cited in this article and wrote the manuscript. EK guided the research and critically revised the manuscript.

Conflicts of Interest

None declared.

References

1. Wenkebach U, Pollwein B, Finsterer U. Visualization of large datasets in intensive care. Proc Annu Symp Comput Appl Med Care 1992:18-22 [FREE Full text] [Medline: 1482864]
2. Wang J, Chen Y, Hua R, Wang P, Fu J. A distributed big data storage and data mining framework for solar-generated electricity quantity forecasting. Proc. SPIE 8333, Photonics and Optoelectronics Meetings (POEM) 2011 [FREE Full text] [doi: 10.1117/12.919640]
3. Wang JZ, Chen YJ, Hua R, Wang P, Fu J. A distributed big data storage and data mining framework for solar-generated electricity quantity forecasting. Proc. SPIE 8333, Photonics and Optoelectronics Meetings (POEM) 2011 2011 [FREE Full text] [doi: 10.1117/12.919640]
4. Fu J, Chen ZH, Wang JC, He MQ, Wang JZ. Distributed storage system big data mining based on HPC application-A solar photovoltaic forecasting system practice. Information-Tokyo 2012;15(9):3749-3755.
5. Brinkmann BH, Bower MR, Stengel KA, Worrell GA, Stead M. Large-scale electrophysiology: acquisition, compression, encryption, and storage of big data. J Neurosci Methods 2009;180(1):185-192 [FREE Full text] [doi: 10.1016/j.jneumeth.2009.03.022] [Medline: 19427545]
6. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, et al. Big data: The future of biocuration. Nature 2008;455(7209):47-50 [FREE Full text] [doi: 10.1038/455047a] [Medline: 18769432]
7. Calatroni A, Roggen D, Troster G. Collection and curation of a large reference dataset for activity recognition. Ieee Sys Man Cybern 2011 2011:30-35. [doi: 10.1109/ICSMC.2011.6083638]

8. O'Driscoll A, Daugelaite J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics. J Biomed Inform 2013;46(5):774-781. [doi: 10.1016/j.jbi.2013.07.001] [Medline: 23872175]
9. Lee KK, Tang WC, Choi KS. Alternatives to relational database: comparison of NoSQL and XML approaches for clinical data storage. Comput Methods Programs Biomed 2013;110(1):99-109. [doi: 10.1016/j.cmpb.2012.10.018] [Medline: 23177219]
10. Beyer MA, Douglas L. The Importance of 'Big Data': A Definition. 2012 URL: http://www.gartner.com/it-glossary/big-data/ [accessed 2013-08-25] [WebCite Cache ID 6J7x1gswQ]
11. Olbers D, Müller P, Willebrand J. Inverse technique analysis of a large data set. Physics of the Earth and Planetary Interiors 1976;12(2-3):248-252. [doi: 10.1016/0031-9201(76)90054-6]
12. Byth DE, Eisemann RL, de Lacy IH. Two-way pattern analysis of a large data set to evaluate genotypic adaptation. Heredity 1976;37(2):215-230. [doi: 10.1038/Hdy.1976.84]
13. Chaudron J, Assenlineau L, Renon H. A new modification of the Redlich—Kwong equation of state based on the analysis of a large set of pure component data. Chemical Engineering Science 1973;28(3):839-846. [doi: 10.1016/0009-2509(77)80018-3]
14. Graefe JF, Wood RW. Dealing with large data sets. Neurotoxicol Teratol 1990;12(5):449-454. [Medline: 2247031]
15. Ackerman MJ. Big data. J Med Pract Manage 2012;28(2):153-154. [Medline: 23167038]
16. Trelles O, Prins P, Snir M, Jansen RC. Big data, but are we ready? Nat Rev Genet 2011 Mar;12(3):224. [doi: 10.1038/nrg2857-c1] [Medline: 21301471]
17. analysis ATK. Big Data and the Creative Destruction of Today's Business Models. 2013 URL: http://www.atkearney.com/strategic-it/ideas-insights/article/-/asset_publisher/LCcgOeS4t85g/content/big-data-and-the-creative-destruction-of-today-s-business-models/10192 [accessed 2013-11-25] [WebCite Cache ID 6LOmyDbvg]
18. Halevi G, Moed HF. The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature. 2012 URL: http://www.researchtrends.com/issue-30-september-2012/the-evolution-of-big-data-as-a-research-and-scientific-topic-overview-of-the-literature/ [accessed 2013-11-25] [WebCite Cache ID 6LOn9qSGu]
19. Hutchins J, Ihler A, Smyth P. Probabilistic analysis of a large-scale urban traffic sensor data set. Knowledge Discovery from Sensor Data 2010;5840:94-114. [doi: 10.1007/978-3-642-12519-5_6]
20. Sleutel S, De Neve S, Beheydt D, Li C, Hofman G. Regional simulation of long-term organic carbon stock changes in cropland soils using the DNDC model: 1. Large-scale model validation against a spatially explicit data set. Soil Use Manage 2006;22(4):342-351. [doi: 10.1111/j.1475-2743.2006.00045.x]
21. Cianchini G, De Santis A, Balasis G, Mandea M, Qamili E. Entropy based analysis of satellite magnetic data for searching possible electromagnetic signatures due to big earthquakes. Ma Comput Sci Eng 2009:29-35 [FREE Full text]
22. Issenberg S. How President Obama's campaign used big data to rally indicidual voters. Technol Rev 2013;116(1):38-49 [FREE Full text]
23. Kessler RC, Brown RL, Broman CL. Sex differences in psychiatric help-seeking: evidence from four large-scale surveys. J Health Soc Behav 1981 Mar;22(1):49-64. [Medline: 7240706]
24. Lewis M. Moneyball: The Art of Winning an Unfair Game: W. In: Moneyball: The Art of Winning an Unfair Game. New York, New York: W. W. Norton & Company; 2003.
25. Ekins S, Williams AJ. When pharmaceutical companies publish large datasets: an abundance of riches or fool's gold? Drug Discov Today 2010;15(19-20):812-815. [doi: 10.1016/j.drudis.2010.08.010] [Medline: 20732447]
26. Grimley Evans J, Tallis RC. A new beginning for care for elderly people? BMJ 2001;322(7290):807-808 [FREE Full text] [Medline: 11290619]
27. Jee K, Kim GH. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. Healthc Inform Res 2013;19(2):79-85 [FREE Full text] [doi: 10.4258/hir.2013.19.2.79] [Medline: 23882412]
28. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013;309(13):1351-1352. [doi: 10.1001/jama.2013.393] [Medline: 23549579]
29. Toga AW. The clinical value of large neuroimaging data sets in Alzheimer's disease. Neuroimaging Clin N Am 2012;22(1):107-118, ix [FREE Full text] [doi: 10.1016/j.nic.2011.11.008] [Medline: 22284737]
30. Bakke PS, Zhu G, Gulsvik A, Kong X, Agusti AG, Calverley PM, et al. Candidate genes for COPD in two large data sets. Eur Respir J 2011;37(2):255-263 [FREE Full text] [doi: 10.1183/09031936.00091709] [Medline: 20562129]
31. Solomon BD, Nguyen AD, Bear KA, Wolfsberg TG. Clinical genomic database. Proc Natl Acad Sci U S A 2013;110(24):9851-9855 [FREE Full text] [doi: 10.1073/pnas.1302575110] [Medline: 23696674]
32. Liu XL, Du JP, Li WZ, Zuo M, Han ZM. Data warehousing for data mining based on Olap technology. In: Ciict 2008: Proceedings of China-Ireland International Conference on Information and Communications Technologies. 2008 Presented at: China-Ireland International Conference on Information and Communications Technologies; 26-28 Sept. 2008; Beijing, China p. 176-179. [doi: 10.1049/cp:20080786]
33. Lohr S. The New York Times. 2012. The Age of Big Data URL: http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=1&_r=0 [WebCite Cache ID 6J7xJAwId]

34. McCann E. EHR Boosts ROI. 2013 URL: http://www.healthcareitnews.com/news/ehr-boost-roi-revenue-medical-group [accessed 2013-11-25] [WebCite Cache ID 6LOnNYQHV]
35. group CIW. EMR Benefits and Return on Investment Categories. 2008 URL: http://www.informatics-review.com/wiki/index.php/EMR_Benefits_and_Return_on_Investment_Categories [accessed 2013-11-25] [WebCite Cache ID 6LOndM19u]
36. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 2001 Dec 4;135(11):982-989. [Medline: 11730399]
37. Mancia G, Omboni S, Ravogli A, Parati G, Zanchetti A. Ambulatory blood pressure monitoring in the evaluation of antihypertensive treatment: additional information from a large data base. Blood Press 1995;4(3):148-156. [Medline: 7670648]
38. Reichelt JG, Heimdal K, Møller P, Dahl AA. BRCA1 testing with definitive results: a prospective study of psychological distress in a large clinic-based sample. Familial Cancer 2002;3(1):21-28. [doi: 10.1023/B:FAME.0000026820.32469.4a]
39. Mitchell PB, Johnston AK, Corry J, Ball JR, Malhi GS. Characteristics of bipolar disorder in an Australian specialist outpatient clinic: comparison across large datasets. Aust N Z J Psychiatry 2009;43(2):109-117. [doi: 10.1080/00048670802607220] [Medline: 19153918]
40. NHGRI. Request for Information (RFI): Input on Development of Analysis MethodsSoftware for Big Data. 2013 URL: http://grants.nih.gov/grants/guide/notice-files/NOT-HG-13-014.html [accessed 2013-08-27] [WebCite Cache ID 6JB6Y33gd]
41. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. Stroke 1991;22(8):983-988 [FREE Full text] [Medline: 1866765]
42. Hubert HB, Feinleib M, McNamara PM, Castelli WP. Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the Framingham Heart Study. Circulation 1983;67(5):968-977. [Medline: 6219830]
43. Gebregziabher M, Egede L, Gilbert GE, Hunt K, Nietert PJ, Mauldin P. Fitting parametric random effects models in very large data sets with application to VHA national data. BMC Med Res Methodol 2012;12:163 [FREE Full text] [doi: 10.1186/1471-2288-12-163] [Medline: 23095325]
44. Bermejo-Sánchez E, Cuevas L, Amar E, Bakker MK, Bianca S, Bianchi F, et al. Amelia: a multi-center descriptive epidemiologic study in a large dataset from the International Clearinghouse for Birth Defects Surveillance and Research, and overview of the literature. Am J Med Genet C Semin Med Genet 2011;157C(4):288-304. [doi: 10.1002/ajmg.c.30319] [Medline: 22002956]
45. Cross SS, Harrison RF. Discriminant histological features in the diagnosis of chronic idiopathic inflammatory bowel disease: analysis of a large dataset by a novel data visualisation technique. J Clin Pathol 2002;55(1):51-57 [FREE Full text] [Medline: 11825925]
46. Kelly D, C Zhang Q, M Soucie J, Manco-Johnson M, Dimichele D, Joint Outcome Subcommittee of the Coordinating Committee for the Universal Data Collection Databasethe Hemophilia Treatment Center Network Investigators. Prevalence of clinical hip abnormalities in haemophilia A and B: an analysis of the UDC database. Haemophilia 2013;19(3):426-431. [doi: 10.1111/hae.12073] [Medline: 23252621]
47. Siregar S, Roes KC, van Straten AH, Bots ML, van der Graaf Y, van Herwerden LA, et al. Statistical methods to monitor risk factors in a clinical database: example of a national cardiac surgery registry. Circ Cardiovasc Qual Outcomes 2013;6(1):110-118. [doi: 10.1161/CIRCOUTCOMES.112.968800] [Medline: 23322806]
48. Elshazly MB, Martin SS, Blaha MJ, Joshi PH, Toth PP, McEvoy JW, et al. Non-high-density lipoprotein cholesterol, guideline targets, and population percentiles for secondary prevention in 1.3 million adults: the VLDL-2 Study (very large database of lipids). J Am Coll Cardiol 2013 Nov 19;62(21):1960-1965. [doi: 10.1016/j.jacc.2013.07.045] [Medline: 23973689]
49. Chan L, McGarey P. Using large datasets for population-based health research. In: Gallin JI, Ognibene FP. eds. Principles and Practice of Clinical Research. 3rd ed. Maryland Heights, MO: Elsevier, Inc; 2012:371-381.
50. Ursum J, Bos WH, van de Stadt RJ, Dijkmans BA, van Schaardenburg D. Different properties of ACPA and IgM-RF derived from a large dataset: further evidence of two distinct autoantibody systems. Arthritis Res Ther 2009;11(3):R75 [FREE Full text] [doi: 10.1186/ar2704] [Medline: 19460147]
51. Ajdacic-Gross V, Vetter S, Müller M, Kawohl W, Frey F, Lupi G, et al. Risk factors for stuttering: a secondary analysis of a large data base. Eur Arch Psychiatry Clin Neurosci 2010;260(4):279-286. [doi: 10.1007/s00406-009-0075-4] [Medline: 19826856]
52. Monnelly EP, Locastro JS, Gagnon D, Young M, Fiore LD. Quetiapine versus trazodone in reducing rehospitalization for alcohol dependence: a large data-base study. J Addict Med 2008;2(3):128-134. [doi: 10.1097/ADM.0b013e318165cb56] [Medline: 21768982]
53. From AM, Bartholmai BJ, Williams AW, Cha SS, Pflueger A, McDonald FS. Sodium bicarbonate is associated with an increased incidence of contrast nephropathy: a retrospective cohort study of 7977 patients at mayo clinic. Clin J Am Soc Nephrol 2008;3(1):10-18 [FREE Full text] [doi: 10.2215/CJN.03100707] [Medline: 18057306]
54. Habash-Bseiso DE, Rokey R, Berger CJ, Weier AW, Chyou PH. Accuracy of noninvasive ejection fraction measurement in a large community-based clinic. Clin Med Res 2005;3(2):75-82 [FREE Full text] [Medline: 16012124]
55. Zhang Y, Hemond MS. Uncovering the predictive value of minimum blood glucose through statistical analysis of a large clinical dataset. AMIA Annu Symp Proc 2009;2009:725-729 [FREE Full text] [Medline: 20351948]

56. Morgan CL, Evans M, Toft AD, Jenkins-Jones S, Poole CD, Currie CJ. Clinical effectiveness of biphasic insulin aspart 30:70 versus biphasic human insulin 30 in UK general clinical practice: a retrospective database study. Clin Ther 2011;33(1):27-35. [doi: 10.1016/j.clinthera.2011.01.023] [Medline: 21397771]
57. Kobayashi H, Miyata H, Gotoh M, Baba H, Kimura W, Kitagawa Y, et al. Risk model for right hemicolectomy based on 19,070 Japanese patients in the National Clinical Database. J Gastroenterol 2013 Jul 27. [doi: 10.1007/s00535-013-0860-8] [Medline: 23892987]
58. International Warfarin Pharmacogenetics Consortium, Klein TE, Altman RB, Eriksson N, Gage BF, Kimmel SE, et al. Estimation of the warfarin dose with clinical and pharmacogenetic data. N Engl J Med 2009;360(8):753-764 [FREE Full text] [doi: 10.1056/NEJMoa0809329] [Medline: 19228618]
59. Moreno-De-Luca D, Sanders SJ, Willsey AJ, Mulle JG, Lowe JK, Geschwind DH, et al. Using large clinical data sets to infer pathogenicity for rare copy number variants in autism cohorts. Mol Psychiatry 2013;18(10):1090-1095 [FREE Full text] [doi: 10.1038/mp.2012.138] [Medline: 23044707]
60. Koefoed P, Andreassen OA, Bennike B, Dam H, Djurovic S, Hansen T, et al. Combinations of SNPs related to signal transduction in bipolar disorder. PLoS One 2011;6(8):e23812 [FREE Full text] [doi: 10.1371/journal.pone.0023812] [Medline: 21897858]
61. Hill B, Proulx J, Zeng-Treitler Q. Exploring the use of large clinical data to inform patients for shared decision making. Stud Health Technol Inform 2013;192:851-855. [Medline: 23920678]
62. Li H, Giger ML, Yuan Y, Chen W, Horsch K, Lan L, et al. Evaluation of computer-aided diagnosis on a large clinical full-field digital mammographic dataset. Acad Radiol 2008;15(11):1437-1445 [FREE Full text] [doi: 10.1016/j.acra.2008.05.004] [Medline: 18995194]
63. Li H, Giger ML, Lan L, Bancroft Brown J, MacMahon A, Mussman M, et al. Computerized analysis of mammographic parenchymal patterns on a large clinical dataset of full-field digital mammograms: robustness study with two high-risk datasets. J Digit Imaging 2012;25(5):591-598 [FREE Full text] [doi: 10.1007/s10278-012-9452-z] [Medline: 22246204]
64. Lin KW, Tharp M, Conway M, Hsieh A, Ross M, Kim J, et al. Feasibility of using Clinical Element Models (CEM) to standardize phenotype variables in the database of genotypes and phenotypes (dbGaP). PLoS One 2013;8(9):e76384 [FREE Full text] [doi: 10.1371/journal.pone.0076384] [Medline: 24058713]
65. Stephen R, Boxwala A, Gertman P. Feasibility of using a large Clinical Data Warehouse to automate the selection of diagnostic cohorts. AMIA Annu Symp Proc 2003:1019 [FREE Full text] [Medline: 14728522]
66. Roelofs E, Persoon L, Nijsten S, Wiessler W, Dekker A, Lambin P. Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. Radiother Oncol 2013;108(1):174-179. [doi: 10.1016/j.radonc.2012.09.019] [Medline: 23394741]
67. Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 2013 Sep 28. [doi: 10.1007/s00439-013-1358-4] [Medline: 24077912]
68. Hu H, Brzeski H, Hutchins J, Ramaraj M, Qu L, Xiong R, et al. Biomedical informatics: development of a comprehensive data warehouse for clinical and genomic breast cancer research. Pharmacogenomics 2004;5(7):933-941. [doi: 10.1517/14622416.5.7.933] [Medline: 15469411]
69. Savas S. A curated database of genetic markers from the angiogenesis/VEGF pathway and their relation to clinical outcome in human cancers. Acta Oncol 2012;51(2):243-246. [doi: 10.3109/0284186X.2011.636758] [Medline: 22150118]
70. Singh SK, Malik A, Firoz A, Jha V. CDKD: a clinical database of kidney diseases. BMC Nephrol 2012;13:23 [FREE Full text] [doi: 10.1186/1471-2369-13-23] [Medline: 22540288]
71. Suzuki H, Gotoh M, Sugihara K, Kitagawa Y, Kimura W, Kondo S, et al. Nationwide survey and establishment of a clinical database for gastrointestinal surgery in Japan: Targeting integration of a cancer registration system and improving the outcome of cancer treatment. Cancer Sci 2011;102(1):226-230. [doi: 10.1111/j.1349-7006.2010.01749.x] [Medline: 20961361]
72. Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. J Gen Intern Med 2013;28(suppl 3):S660-S665. [doi: 10.1007/s11606-013-2455-8] [Medline: 23797912]
73. Han Y, Itälä T, Hämäläinen M. Citizen Centric Architecture approach - taking e-health forward by integrating citizens and service providers. Stud Health Technol Inform 2010;160(Pt 2):907-911. [Medline: 20841816]
74. Boussadi A, Caruba T, Zapletal E, Sabatier B, Durieux P, Degoulet P. A clinical data warehouse-based process for refining medication orders alerts. J Am Med Inform Assoc 2012;19(5):782-785 [FREE Full text] [doi: 10.1136/amiajnl-2012-000850] [Medline: 22523345]
75. Hernandez P, Podchiyska T, Weber S, Ferris T, Lowe H. AMIA Annu Symp Proc. 2009. Automated mapping of pharmacy orders from two electronic health record systems to RxNorm within the STRIDE clinical data warehouse URL: http://europepmc.org/abstract/MED/20351858/reload=0;jsessionid=4aLVRuWM5ugKTsb134ug.0 [accessed 2014-01-09] [WebCite Cache ID 6MVbvtUxj]
76. Cuggia M, Garcelon N, Campillo-Gimenez B, Bernicot T, Laurent JF, Garin E, et al. Roogle: an information retrieval engine for clinical data warehouse. Stud Health Technol Inform 2011;169:584-588. [Medline: 21893816]

77. Zhou X, Chen S, Liu B, Zhang R, Wang Y, Li P, et al. Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. Artif Intell Med 2010;48(2-3):139-152. [doi: 10.1016/j.artmed.2009.07.012] [Medline: 20122820]
78. Zapletal E, Rodon N, Grabar N, Degoulet P. Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case. Stud Health Technol Inform 2010;160(Pt 1):193-197. [Medline: 20841676]
79. Hanss S, Schaaf T, Wetzel T, Hahn C, Schrader T, Tolxdorff T. Integration of decentralized clinical data in a data warehouse: a service-oriented design and realization. Methods Inf Med 2009;48(5):414-418. [doi: 10.3414/ME9240] [Medline: 19657544]
80. Evans RS, Lloyd JF, Pierce LA. AMIA Annu Symp Proc. 2012. Clinical use of an enterprise data warehouse URL: http://europepmc.org/abstract/MED/23304288/reload=0;jsessionid=P5NVyPK8BYs9OV0Ra8hg.0 [accessed 2014-01-09] [WebCite Cache ID 6MVc6BqLh]
81. Harrison JH. Introduction to the mining of clinical data. Clin Lab Med 2008;28(1):1-7. [doi: 10.1016/j.cll.2007.10.001] [Medline: 18194715]
82. Phan JH, Quo CF, Cheng C, Wang MD. Multiscale integration of -omic, imaging, and clinical data in biomedical informatics. IEEE Rev Biomed Eng 2012;5:74-87. [doi: 10.1109/RBME.2012.2212427] [Medline: 23231990]
83. Sinha A, Hripcsak G, Markatou M. Large datasets in biomedicine: a discussion of salient analytic issues. J Am Med Inform Assoc 2009;16(6):759-767 [FREE Full text] [doi: 10.1197/jamia.M2780] [Medline: 19717808]
84. Drai D, Grodzinsky Y. A new empirical angle on the variability debate: quantitative neurosyntactic analyses of a large data set from Broca's aphasia. Brain Lang 2006;96(2):117-128. [doi: 10.1016/j.bandl.2004.10.016] [Medline: 16115671]
85. Bellman R, Kalaba R. DYNAMIC PROGRAMMING AND STATISTICAL COMMUNICATION THEORY. Proc Natl Acad Sci U S A 1957;43(8):749-751 [FREE Full text] [Medline: 16590080]
86. Borthakur D. The Hadoop Distributed File System: Architecture and Design. 2007 URL: https://hadoop.apache.org/docs/r0.18.0/hdfs_design.pdf [accessed 2013-11-25] [WebCite Cache ID 6LOoNyUXR]
87. Chang F, Dean J, Ghemawat S. Bigtable: A Distributed Storage System for Structured Data. 2006 Presented at: the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI'06); 2006; Seattle, WA, US p. 205-218.
88. Litt S. NoSQL: The Unix Database (With awk). 2007 URL: http://www.troubleshooters.com/lpm/200704/200704.htm [accessed 2013-11-25] [WebCite Cache ID 6LOp7PTH6]
89. van der Burgt YE, Taban IM, Konijnenburg M, Biskup M, Duursma MC, Heeren RM, et al. Parallel processing of large datasets from NanoLC-FTICR-MS measurements. J Am Soc Mass Spectrom 2007;18(1):152-161. [doi: 10.1016/j.jasms.2006.09.005] [Medline: 17055738]
90. Stata Corporation. Stata reference manual : release 6. In: Stata Reference Manual Set, 4vol: Release 6. College Station, TX: Stata Corp; 1999.
91. R Development Core Team. official website for R. Vienna, Austria The R Project for Statistical Computing URL: http://www.r-project.org/ [accessed 2014-01-08] [WebCite Cache ID 6MU3hBSqT]
92. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19(2):185-193 [FREE Full text] [Medline: 12538238]
93. Pleil JD, Stiegel MA, Madden MC, Sobus JR. Heat map visualization of complex environmental and biomarker measurements. Chemosphere 2011;84(5):716-723. [doi: 10.1016/j.chemosphere.2011.03.017] [Medline: 21492901]
94. Garcia-Betancur JC, Menendez MC, Del Portillo P, Garcia MJ. Alignment of multiple complete genomes suggests that gene rearrangements may contribute towards the speciation of Mycobacteria. Infect Genet Evol 2012;12(4):819-826. [doi: 10.1016/j.meegid.2011.09.024] [Medline: 22008279]
95. Aita T, Nishigaki K. A visualization of 3D proteome universe: mapping of a proteome ensemble into 3D space based on the protein-structure composition. Mol Phylogenet Evol 2011;61(2):484-494. [doi: 10.1016/j.ympev.2011.06.020] [Medline: 21762784]
96. Lenzerini M. Data Integration: A Theoretical Perspective. 2002 Presented at: Proceedings of the ACM Symposium on Principles of Database Systems (PODS); 2002; Roma, Italy p. 233-246. [doi: 10.1145/543613.543644]
97. Halevy A, Rajaraman A, Ordille J. Data integration: the teenage years. 2006 Presented at: VLDB '06 Proceedings of the 32nd international conference on Very large data bases; 2006; Seoul, Korea p. 9-16.
98. Haas LM, Lin ET, Roth MA. Data integration through database federation. IBM Syst. J 2002;41(4):578-596. [doi: 10.1147/sj.414.0578]
99. Shyu C, Ytreberg FM. Reducing the bias and uncertainty of free energy estimates by using regression to fit thermodynamic integration data. J Comput Chem 2009;30(14):2297-2304. [doi: 10.1002/jcc.21231] [Medline: 19266482]
100. Taylor W, Gladman D, Helliwell P, Marchesoni A, Mease P, Mielants H, CASPAR Study Group. Classification criteria for psoriatic arthritis: development of new criteria from a large international study. Arthritis Rheum 2006;54(8):2665-2673 [FREE Full text] [doi: 10.1002/art.21972] [Medline: 16871531]
101. Gschwind R, Robert Y. Analyse der zeitlichen Veränderungen der Papillen-Reflexion mittels Hauptkomponentenanalyse. Klin Monatsbl Augenheilkd 2008;190(04):249. [doi: 10.1055/s-2008-1050370]
102. Santafé G, Lozano JA, Larrañaga P. Bayesian model averaging of naive Bayes for clustering. IEEE Trans Syst Man Cybern B Cybern 2006;36(5):1149-1161. [Medline: 17036820]

103. Farré J, Cabrera JA, Romero J, Rubio JM. Therapeutic decision tree for patients with sustained ventricular tachyarrhythmias or aborted cardiac arrest: a critical review of the Antiarrhythmics Versus Implantable Defibrillator trial and the Canadian Implantable Defibrillator Study. Am J Cardiol 2000;86(9A):44K-51K. [Medline: 11084100]
104. Lisboa PJ. A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Netw 2002;15(1):11-39. [Medline: 11958484]
105. Chen HF. In silico log P prediction for a large data set with support vector machines, radial basis neural networks and multiple linear regression. Chem Biol Drug Des 2009;74(2):142-147. [doi: 10.1111/j.1747-0285.2009.00840.x] [Medline: 19549084]
106. Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. Am J Public Health 1996;86(5):726-728. [doi: 10.2105/Ajph.86.5.726]
107. Bender R, Lange S. Adjusting for multiple testing—when and how? Journal of Clinical Epidemiology 2001;54(4):343-349. [doi: 10.1016/S0895-4356(00)00314-0]
108. Broberg P. A comparative review of estimates of the proportion unchanged genes and the false discovery rate. BMC Bioinformatics 2005;6:199 [FREE Full text] [doi: 10.1186/1471-2105-6-199] [Medline: 16086831]
109. van der Laan MJ, Dudoit S, Pollard KS. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. Stat Appl Genet Mol Biol 2004;3. [doi: 10.2202/1544-6115.1041]
110. Sanders CM, Saltzstein SL, Schultzel MM, Nguyen DH, Stafford HS, Sadler GR. Understanding the limits of large datasets. J Cancer Educ 2012;27(4):664-669. [doi: 10.1007/s13187-012-0383-7] [Medline: 22729362]
111. Derman E. Models. In: Models.Behaving.Badly. Why Confusing Illusion with Reality Can Lead to Disaster, on Wall Street and in Life. New York, New York: Free Press; 2012.
112. Kobayashi T, Kishimoto M, Swearingen CJ, Filopoulos MT, Ohara Y, Tokuda Y, et al. Differences in clinical manifestations, treatment, and concordance rates with two major sets of criteria for Behçet's syndrome for patients in the US and Japan: data from a large, three-center cohort study. Mod Rheumatol 2013;23(3):547-553. [doi: 10.1007/s10165-012-0696-8] [Medline: 22752504]
113. Jacobs A. The pathologies of big data. Commun ACM 2009;52(8):36. [doi: 10.1145/1536616.1536632]

Abbreviations

BD2K: Big Data to Knowledge
CS: computer science
SNP: single-nucleotide polymorphism
ZB: zettabytes

Edited by G Eysenbach; submitted 27.08.13; peer-reviewed by L Toldo, J Gao; comments to author 27.10.13; revised version received 25.11.13; accepted 08.12.13; published 17.01.14.

Please cite as:
Wang W, Krishnan E
Big Data and Clinicians: A Review on the State of the Science
JMIR Med Inform 2014;2(1):e1
URL: http://www.medinform.jmir.org/2014/1/e1/
doi:10.2196/medinform.2913
PMID:25600256

Original Paper

Increased Workload for Systematic Review Literature Searches of Diagnostic Tests Compared With Treatments: Challenges and Opportunities

Henry Petersen[1*], MPhil; Josiah Poon[1*], PhD; Simon K Poon[1*], MPH, PhD; Clement Loy[2*], FRACP, PhD

[1]School of Information Technologies, Faculty of Engineering and IT, University of Sydney, Sydney, Australia

[2]School of Public Health, Faculty of Medicine, University of Sydney, Sydney, Australia

[*]all authors contributed equally

Corresponding Author:
Josiah Poon, PhD
School of Information Technologies
Faculty of Engineering and IT
University of Sydney
School of Information Technologies Building, J12
University of Sydney
Sydney,
Australia
Phone: 61 02 9351 7185
Fax: 61 02 9351 3838
Email: josiah.poon@sydney.edu.au

Abstract

Background: Comprehensive literature searches are conducted over multiple medical databases in order to meet stringent quality standards for systematic reviews. These searches are often very laborious, with authors often manually screening thousands of articles. Information retrieval (IR) techniques have proven increasingly effective in improving the efficiency of this process. IR challenges for systematic reviews involve building classifiers using training data with very high class-imbalance, and meeting the requirement for near perfect recall on relevant studies. Traditionally, most systematic reviews have focused on questions relating to treatment. The last decade has seen a large increase in the number of systematic reviews of diagnostic test accuracy (DTA).

Objective: We aim to demonstrate that DTA reviews comprise an especially challenging subclass of systematic reviews with respect to the workload required for literature screening. We identify specific challenges for the application of IR to literature screening for DTA reviews, and identify potential directions for future research.

Methods: We hypothesize that IR for DTA reviews face three additional challenges, compared to systematic reviews of treatments. These include an increased class-imbalance, a broader definition of the target class, and relative inadequacy of available metadata (ie, medical subject headings (MeSH) terms for medical literature analysis and retrieval system online). Assuming these hypotheses to be true, we identify five manifestations when we compare literature searches of DTA versus treatment. These manifestations include: an increase in the average number of articles screened, and increase in the average number of full-text articles obtained, a decrease in the number of included studies as a percentage of full-text articles screened, a decrease in the number of included studies as a percentage of all articles screened, and a decrease in the number of full-text articles obtained as a percentage of all articles screened. As of July 12 2013, 13 published Cochrane DTA reviews were available and all were included. For each DTA review, we randomly selected 15 treatment reviews published by the corresponding Cochrane Review Group (N=195). We then statistically tested differences in these five hypotheses, for the DTA versus treatment reviews.

Results: Despite low statistical power caused by the small sample size for DTA reviews, strong (P<.01) or very strong (P<.001) evidence was obtained to support three of the five expected manifestations, with evidence for at least one manifestation of each hypothesis. The observed difference in effect sizes are substantial, demonstrating the practical difference in reviewer workload.

Conclusions: Reviewer workload (volume of citations screened) when screening literature for systematic reviews of DTA is especially high. This corresponds to greater rates of class-imbalance when training classifiers for automating literature screening

for DTA reviews. Addressing concerns such as lower quality metadata and effectively modelling the broader target class could help to alleviate such challenges, providing possible directions for future research.

(JMIR Med Inform 2014;2(1):e11) doi:10.2196/medinform.3037

KEYWORDS

meta-analysis; data mining; review literature; information storage and retrieval; classification and clustering

Introduction

Background

Systematic reviews are a key component in evidence-based medicine and are widely regarded as the highest form of medical evidence [1]. A number of organizations such as the Cochrane collaboration exist to facilitate the generation and dissemination of systematic reviews for a range of clinical questions and fields. For example, Cochrane maintains the Cochrane database of systematic reviews; an extensive database which, at the end of the year 2013 contained over 5000 reviews. Traditionally, systematic reviews have focused on questions related to medical interventions, however recently there has been increasing demand for reviews from other areas (ie, etiology, diagnosis, prognosis, etc). In particular, there has been a substantial increase in demand for reviews of diagnostic test accuracy (DTA) leading to the formation of the Cochrane diagnostic test accuracy working group in 2003.

The high potential cost of omitting relevant studies from medical decision making is well established [2]. In order to meet the stringent recall requirements for systematic reviews, authors must conduct highly sensitive, detailed literature searches. To minimize the possibility of error, these searches in most cases are manually conducted and are eventually time consuming [1]. It is not unusual for an individual review to be conducted over the course of months or even years [3]. As the demand for systematic reviews increases, it is apparent that methods to automate or expedite the review process are essential [4].

In recent times there has been much interest expressed by the information retrieval (IR) community on increasing the automation of literature searches for systematic reviews [5-7]. This automation process typically involves a set of labelled training instances (articles marked as relevant or irrelevant to the target review), and a classification algorithm which is run on these instances to “train” a mapping function (“classifier”) from instances to labels. From the perspective of training such a classifier, systematic reviews present several challenges: the training data is highly imbalanced (ie, the number of included studies will be small as a percentage of all training examples) [5], there is a need for near perfect recall, and it is not clear how to best incorporate partial automation into the systematic review process. Despite the above concerns, these methods have met with limited success. Thus further improvements on the methodology is a clear mandate [8,9].

While the medical community has noted a number of challenges facing authors of DTA reviews [10], there has been no analysis on the differences between reviews of DTA and interventions as an IR problem. For the purposes of this study we consider the term “diagnostic test accuracy” to be defined as broadly as possible (we do not limit ourselves to any particular field or study design and consider a DTA review to be any review evaluating the accuracy of a specific diagnostic test). From an IR perspective one of the key challenges in retrieval for systematic reviews is the level of class-imbalance. We identify DTA reviews as a subclass with particularly high class-imbalance rates through a statistical analysis of the reported literature searches from a number of Cochrane reviews of DTA and treatment. Our analysis also identifies two potential causes, which from an IR standpoint provide potential starting points in reducing the additional level of class-imbalance.

The remainder of this section briefly describes the literature search process for systematic reviews and previous applications of IR to the systematic review process. For the sake of brevity, only prior work relating to IR challenges from literature searches where differences between DTA and interventions exist are covered. The interested reader is directed to other literature for more information [11,12].

Overview of Systematic Reviews

While the exact process for conducting a systematic review varies according to the type of clinical question (ie, diagnosis, intervention, etiology), all systematic reviews can be said to follow several steps [13]. These include question and inclusion criteria formulation, literature search, literature screening, quality assessment, and data synthesis, analysis and interpretation.

For brevity’s sake a summary of the entire systematic review process is not presented. Instead we include a brief summary of the first three stages. For further information the interested reader is directed to literature such as Wright et al [1] or the Cochrane handbooks for reviews of interventions [14] and DTA [15].

Question and Inclusion Criteria Formulation

Systematic reviews begin with the formulation of a highly specific research question and associated inclusion criteria. Inclusion criteria for Cochrane Systematic Reviews are formulated according to specific concepts that depend on the type of clinical question being answered. For example, in Cochrane Reviews of diagnostic test accuracy, separate criteria are formulated for the type of study, index and comparator tests, target condition and the desired reference standard [15]. A similar set of criteria (referred to as the PICO criteria—Population, Intervention, Comparison, Outcome) exists for questions related to interventions [14].

Literature Search

Review authors will then query multiple databases to identify potential relevant studies (usually medical literature analysis and retrieval system online (MEDLINE) and excerpta medica

database (EMBASE) at a minimum although other resources do exist). In order to facilitate this process, citations indexed in these databases are typically annotated with entries from a controlled hierarchy of medical concepts that can be used for search and retrieval (examples include the MeSH for MEDLINE or EMTREE for EMBASE).

Literature searches for Cochrane Systematic Reviews are typically conducted by identifying references containing relevant MeSH and free text terms. Cochrane Reviews of interventions usually identify multiple MeSH terms relating to several key concepts of the review. Searches for each of these concepts are run using the identified MeSH terms, with the union of the search results selected for further screening. Literature searches for systematic reviews of diagnostic test accuracy are similar, however the methodological search filter is often omitted [16]. While much research has been done on developing highly sensitive DTA filters [17-22], the broader community has yet to develop a consensus on their use in DTA reviews (for example the Cochrane handbook for DTA reviews recommends against the "routine use of methodology search filters"[15]).

Literature Screening

References returned by the literature search are manually compared against the inclusion criteria for the reviews in a two stage process. Initially, two reviewers independently screen title and abstracts for all references, with full-text articles obtained for any potentially relevant citations. These full-text articles are then screened again by both reviewers.

To meet the requirement for near perfect recall, the number of references screened can often be many times greater, often one or two orders of magnitude than the number included in the final review. Karimi et al noted that when screening citations, each individual document may require several minutes to process [23]. It is apparent that even small reductions in the number of citations screened could result in a significant reduction in reviewer time and effort. Still, the high rates of class-imbalance, combined with the stringent recall requirements present a significant IR challenge.

State of the Art

A major concern for IR with systematic reviews is dealing with highly imbalanced training data when building classifiers (ie, the number of available examples of relevant articles for a given review will be small relative to the number of irrelevant ones, leading to models which can be biased towards the non-relevant studies). Addressing this class-imbalance has been a key feature of much of the relevant IR literature [5]. Existing techniques have met with some success, however improvements in performance are still required, especially for those with higher rates of imbalance [8,9].

In addition to high levels of class-imbalance, IR for systematic reviews must also meet stringent recall requirements. In other words, there is a large difference in the cost of false positive and false negative errors for IR algorithms when identifying citations for inclusion into systematic reviews. Prior work addressing this issue include the modified voting perceptron method of Cohen et al [24], the factorized Complement Naïve Bayes model of Matwin et al [25], and support vector machine based approaches by Cohen et al [6,26] and Wallace et al [7,27]

Attention has also been directed towards the best approaches on combining IR techniques with the systematic review process. Frunza et al [5] describe an approach based on having authors manually screen some percentage of all citations, which are then used as training data to build a classifier to be run on the remaining articles. In contrast, Wallace et al [7, 27, 28] describe an active learning approach, where the classifier is built in an iterative process. Here the algorithm particularly selects those citations for which manual annotation would provide the greatest improvement. Finally, work exists addressing the similar task of identifying studies to update existing reviews [24,29]. Automation of the review update task is similar to classification for the initial review, however it fits much better with the traditional classification model in which separate training and test sets are used (ie, annotations from the original search can be used to train the classifier for the update task).

There has also been some interest in applying classification to assign relevant MeSH terms to citations from MEDLINE [30], as well as retrieval of studies of high methodological quality [31]. For example, on employing articles retrieved from the American College of Physicians (ACP) journal club as training data, Aphinyanaphongs et al [31] evaluated a range of common algorithms and reported that their preliminary results showed good performance on identifying high quality DTA studies.

While such results may at first seem to contradict the difficulty of creating high quality DTA classifiers, the distinction between general retrieval of DTA studies and retrieval of 'high quality' DTA studies should be noted. Aphinyanaphongs et al trained their classifier based on citations retrieved from the ACP journal clubs meta-publication which applies strict quality criteria to determine if a citation should be included [31]. As the ACP restrict inclusion to high quality articles this could be expected to significantly reduce variance across the target class, reducing the complexity of the task for any prospective classifier.

Methods

Overview

This section outlines three hypotheses regarding technical challenges faced by both authors and IR researches for DTA reviews. These hypotheses relate to differences in the literature search process between systematic reviews of DTA and treatment. Hypothesis A relates to the screening process as a whole, while Hypotheses B and C relate to stage 2 and stage 1 screening respectively. We describe one or more expected manifestations for each hypothesis. The analysis in this paper reports whether or not the expected manifestations can be observed and if the observations are statistically significant. A tabular summary linking each hypothesis, manifestation, and screening stage is presented in Table 1.

Hypothesis A: Increased Workload for DTA Reviews

A major practical issue when conducting systematic reviews is the workload generated by the volume of citations needing to be screened. Most IR research for systematic reviews has focused specifically on how to deal with the very high rates of

class-imbalance caused by this volume of data. Substantial progress has been made, however it can by no means be considered a solved problem.

This article claims that the number of citations to be screened at each stage of the literature search process is higher for DTA reviews than for those of the treatment. From an IR perspective, this increases the already large class-imbalance between the number of included and excluded studies, thereby again increasing the difficulty of what was already very challenging. Assuming this to be true, one could then expect the following manifestations (restated in Table 1): First, the mean number of search results to be screened will be higher for DTA reviews than for those of treatment. Second, the mean number of full-text articles to be screened will be higher for DTA reviews than for those of treatment. Finally, the number of included studies as a percentage of the number of full-text articles screened would be lower for DTA reviews than for treatment.

Hypothesis B: Increased Target Class Heterogeneity for DTA

The relative heterogeneity of what exactly constitutes a DTA study can be problematic when screening literature for DTA reviews. Quoting from Whiting et al [19], diagnostic test accuracy studies "are heterogeneous, exploring a range of diagnostic techniques and strategies, and are likely to have been conducted using a variety of methods". In addition, there are examples (such as some cohort studies) where one could derive sensitivity and specificity despite the authors not having explicitly calculated them. The ideal DTA filter should be highly sensitive and would include studies such as these.

Our paper suggests that due to this increased difficulty, the percentage of irrelevant citations that cannot be identified on title and abstract alone will be larger for DTA reviews than for treatment. Assuming this to be true, we can expect the following manifestations (restated in Table 1): The mean number of full-text articles to be screened will be higher for DTA reviews than for those of treatment, and the number of included studies as a percentage of the number of full-text articles screened would be lower for DTA reviews than for treatment.

Intuitively, if a given study type is more challenging to identify than another, it can be expected that an author would need to expend greater effort on discerning similar studies. This increased effort could take the form of additional time to screen individual citations, or screening more citations in greater detail (ie, examining the full-text article). Due to the high cost of false negative classifications, it is reasonable to assume that any ambiguity in the initial screening stage would be resolved by obtaining the full-text article rather than putting more effort on the title and abstract. As such, assuming DTA studies to be inherently more challenging to identify than randomized controlled trials, we would expect to observe more full-text articles being screened when conducting DTA reviews.

Hypothesis C: Decreased Suitability of Metadata for DTA

Appropriate use of high quality metadata (ie, MeSH terms for MEDLINE) in literature searches is crucial to generate a manageable number of citations while still remaining confident that no relevant ones would be omitted. It is common to identify thousands of citations at this stage. It follows that as the quality of the available metadata decreases, the total number of citations one would need to screen to maintain this confidence would increase.

It has been noted within the literature that the metadata in many medical databases are more suited to describing concepts related to treatment as opposed to diagnosis [15]. For example, while high quality MeSH terms exist for study types such as randomized controlled trials, the same cannot be said for studies of diagnostic test accuracy. From Whiting et al [19]: "Although MEDLINE includes a number of medical subject headings (MeSH) that capture outcome measures used in test accuracy studies (eg, sensitivity and specificity), these terms are not specific to test accuracy studies and are inconsistently applied by indexers".

This article claims that the quality of metadata is typically lower for DTA reviews than for treatment. Therefore we can expect the following manifestations in literature searches for systematic reviews (restated in Table 1): The mean number of search results to be screened will be higher for DTA reviews than for those of treatment, and the number of full-text articles retrieved as a percentage of the total search results would be lower for DTA reviews than for treatment.

Data Collection

We have identified five expected manifestations of the stated hypotheses on the literature searches for DTA reviews (restated in Table 1). In order to test these claims, summaries of the literature search and screening stages were extracted from a sample of DTA and treatment reviews. Data collected included the number of references retrieved by the original search (SR), the number of references for which full-text papers were screened (FT), the number of references included in the final meta-analysis (INC), and the paired ratios between each of the collected statistics.

Table 1. List of expected manifestations (differences between DTA and treatment reviews) for all hypotheses.

Manifestation	Description	Hypothesis A: Increased Workload	Hypothesis B: Increased Target Class Heterogeneity	Hypothesis C: Decreased Suitability of Metadata
FT[a]	The mean number of full-text articles screened would be higher for DTA reviews than for treatment	Yes	-	-
SR[b]	The mean number of search results would be higher for DTA reviews than for treatment	Yes	-	Yes
INC[c]/ FT	The number of included studies as a percentage of the number of full-text articles screened would be lower for DTA reviews than for treatment	-	Yes	-
INC/SR	The number of included studies as a percentage of the total number of search results would be lower for DTA reviews than for treatment	Yes	-	-
FT/SR	The number of full-text articles retrieved as a percentage of the total search results would be lower for DTA reviews than for treatment	-	-	Yes

[a]number of references for which full-text papers were screened

[b]number of references retrieved by the original search

[c]number of references included in the final meta-analysis

Systematic reviews can be conducted and reported according to varying standards of rigor. This could be problematic for the purposes of our evaluation, as ideally the variation between two samples should be restricted to one review type (ie, DTA or treatment). For systematic reviews published by the Cochrane collaboration, authors are required to follow strict guidelines outlined in the Cochrane handbooks for treatment and DTA reviews [14,15]. Reviews published by Cochrane are widely regarded as meeting very high procedural and reporting standards, and their published guidelines for reviews of DTA and treatment contain a number of shared protocols. As we wish to restrict differences between the samples to whether the reviews are of treatment or DTA, the analysis reported in this paper is performed exclusively on a subset of the Cochrane database.

As of the search date (July 12 2013), Cochrane had published 13 complete systematic reviews of DTA (one from each of the acute respiratory infections [ARI], airways, back, bone, joint, and muscle trauma [BJMT], eyes and vision, gynecological cancer, pregnancy, renal, and stroke Cochrane review groups [CRG], two from the infectious diseases CRG, and three from the Back CRG). A copy of each DTA review was obtained. For each DTA review, 15 non-withdrawn treatment reviews were selected at random from those published by the corresponding CRG. Stratifying the data in this way was intended to account for any variation in search procedures across CRGs, as well as the availability of data within each field generally. A summary of the number of selected treatment reviews for each CRG is presented in Table 2. A list of each selected diagnostic and treatment review is included in the Multimedia Appendix 1. One author then manually collected the desired statistics from the values reported in the literature search summary from each review.

It is important to recall that depending on the specific conditions of each review (DTA or treatment) changes in the search process may be made to find the desired balance between search sensitivity and reviewer workload. Using the values reported by the reviewers (as opposed to manually re-running searches, possibly with the inclusion of more or less sensitive filters) had the added benefit of taking into account the review authors conclusions for the specific domain of each review.

Table 2. Summary of the total number of DTA and treatment reviews randomly selected for inclusion in our analysis, ordered by CRG.

Cochrane review groups	DTA reviews	Treatment reviews
Acute respiratory infections	1	15
Airways	1	15
Back	3	45
Bone, joint, and muscle trauma	1	15
Eyes and vision	1	15
Gynecological cancer	1	15
Infectious diseases	2	30
Pregnancy	1	15
Renal	1	15
Stroke	1	15
Total	13	195

Not all reviews reported the number of citations obtained at each stage of the literature search (eg, some would report only the number of included and full-text articles). Where values were missing or unclear, we made an attempt to contact the review authors by email. If no data could be obtained, a blank value was recorded and the review would be omitted from analyses involving the missing statistical data. For computational reasons, extracted values equal to 0 were also omitted. A summary of the number of extracted values for all data types is given in Table 3. For example, of the 195 randomly selected treatment reviews, the number of full-text articles examined could not be obtained from 62 reviews, hence the number of collected data points for the number of full-text articles in treatment reviews is 133 (as reported in row 2 of Table 3).

Table 3. Table 3. Summary of the sample sizes (number of reviews reporting nonzero values) for evaluating each of the expected manifestations.

	DTA	Treatment
$DATA_{INC}$	13/13	186/195
$DATA_{FT}$	12/13	133/195
$DATA_{SR}$	13/13	101/195
$DATA_{INC / FT}$	12/13	126/195
$DATA_{INC / SR}$	13/13	95/195
$DATA_{FT / SR}$	12/13	92/195

Analysis

Based on prior experience, we expected that the number of reported studies for the literature searches would be heavily skewed. This expectation is supported by comparing the mean and median values for each statistics from the collected treatment reviews (see Table 4); for 5 out of the 6 statistics the mean is approximately twice the value of the median. For example, the number of reported search results collected includes a number of values describing unusually large literature searches. Such values significantly affect the skewedness of the collected data, substantially increasing the mean without affecting the median.

Table 4. Ratio between mean and median for collected treatment reviews.

	Mean	Median	Mean / Median
$DATA_{INC}$	19.56	11.0	1.78
$DATA_{FT}$	71.89	33.00	2.18
$DATA_{SR}$	1799.04	900.00	2.00
$DATA_{INC / FT}$	0.394	0.357	1.11
$DATA_{INC / SR}$	0.033	0.013	2.47
$DATA_{FT / SR}$	0.099	0.046	2.13

In order to compensate for the level of skewness, all reported statistical comparisons are performed using an unequal variance *t* test on ranked data (ie, as an approximation to a non-parametric test); each individual data point is replaced by its index in the sorted set of data. If multiple data points shared a common value the ranked values were averaged. Summaries of the unranked and ranked data are presented in Table 5 and Table 6.

To further illustrate the ranking process, the mean number of search results obtained (as reported in Table 5) was 5144.23 for DTA reviews and 1799.04 for treatment reviews. When the 13 DTA and 101 treatment data points were combined and sorted however, the mean position for DTA reviews was 85.54 and that for the treatment reviews was 52.76 (as reported in Table 6).

Table 5. Summary of mean values for collected statistics.

	$Mean_{DTA}$	$Mean_{Treat}$	$Mean_{DTA}$/ $Mean_{Treat}$
$DATA_{FT}$	191.92 (n=13,s=233.51)	71.89 (n=133,s=154.76)	2.67
$DATA_{SR}$	5144.23 (n=13,s=4109.78)	1799.04 (n=101,s=2530.11)	2.86
$DATA_{INC\ /\ FT}$	0.191 (n=13,s=0.11)	0.394 (n=126,s=0.24)	0.49
$DATA_{INC\ /\ SR}$	0.021 (n=13,s=0.036)	0.033 (n=95,s=0.049)	0.63
$DATA_{FT\ /\ SR}$	0.087 (n=13,s=0.124)	0.100 (n=92,s=0.156)	0.87

Table 6. Summary of ranked data for collected statistics.

	$Mean_{DTA}$	$Median_{DTA}$	$Mean_{Treat}$	$Median_{Treat}$
$DATA_{FT}$	110.67 (n=12,s=27.64)	113.0	68.51 (n=133,s=41.16)	67.0
$DATA_{SR}$	85.54 (n=13,s=27.84)	94.0	52.76 (n=101,s=31.62)	52.0
$DATA_{INC\ /\ FT}$	35.67 (n=12,s=24.69)	29.0	71.63 (n=126,s=39.60)	73.5
$DATA_{INC\ /\ SR}$	40.54 (n=13,s=31.12)	35.0	55.27 (n=95,s=30.76)	56.0
$DATA_{FT\ /\ SR}$	47.5 (n=12,s=30.18)	45.5	52.02 (n=92,s=29.97)	53.5

Results

Overview

The results section is divided into one section for each of the three proposed hypotheses. Summaries of each hypothesis, along with the expected and observed manifestations is presented in Table 7.

Table 7. Summary linking each hypothesis, expected manifestation, and literature screening stage.

	Hypothesis A: Increased workload	Hypothesis B: Increased target class heterogeneity	Hypothesis C: Decreased suitability of metadata
Total articles screened	Increase	-	Increase
	$5144.2_{DTA} > 1799.0_{TR}$		$5144.2_{DTA} > 1799.0_{TR}$
	(*P*=.002)		(*P*=.002)
Full-text articles obtained	Increase	-	Decreased as a % of total articles screened
	$191.9_{DTA} > 71.9_{TR}$		$0.087_{DTA} < 0.100_{TR}$
	(*P*<.001)		(*P*=.65)
Included Articles	Decrease as a % of total articles screened	Decreased as a % of full-text articles obtained	-
	$0.021_{DTA} < 0.033_{TR}$	$0.191_{DTA} < 0.394_{TR}$	
	(*P*=.14)	(*P*<.001)	

Hypothesis A: Increased Workload for DTA Reviews

Comparing the mean absolute number of the search results obtained we observe a 186% increase for reviews of DTA when compared to reviews of interventions (5144.2 vs 1799.0). There was strong evidence that this difference was statistically significant (unequal variance *t* test on ranked data, *P*=.002). Similarly for the mean number of full-text articles obtained we

can observe an increase of 167% (191.9 vs 71.9). Again, there was very strong evidence that this difference was statistically significant (unequal variance *t* test on ranked data, *P*<.001).

We note not only the statistically significant difference in means, but also the substantial difference in effect size. The magnitude of the difference supports the claim that identification of relevant papers is noticeably more complicated for DTA reviews than for those of treatment, and also that there is an increase in difficulty both for authors and any prospective IR system.

Considering the number of included studies as a proportion of the total search results, a decrease of approximately 35% is observed for DTA reviews when compared to reviews of treatment (0.021 vs 0.033). However, despite the large magnitude of the difference there is insufficient evidence to claim statistically significance (unequal variance *t* test on ranked data, *P*=.14). However, the authors urge caution in concluding that no difference exists (see discussion).

Hypothesis B: Increased Target Class Heterogeneity for DTA

Comparing the number of included studies as a percentage of full-text articles examined, an increase of approximately 106% is observed for DTA reviews when compared to those for treatment (0.191 vs 0.394). Very strong evidence was obtained that this difference was significant (unequal variance *t* test on ranked data, *P*<.001).

Again, we note the substantial difference in the observed effect size here. Its magnitude indicates the increased practical difficulty of screening a potentially relevant article for inclusion in a DTA review.

Hypothesis C: Decreased Suitability of Metadata for DTA

As stated in the results section for Hypothesis A, strong evidence was obtained to support an increase in the mean absolute number of search results obtained when comparing reviews of DTA and treatment (unequal variance *t* test on ranked data, *P*=.002). When looking at the number of full-text articles retrieved as a percentage of total search results, one can observe a decrease of approximately 13% for DTA reviews when compared to treatment reviews (0.087 vs 0.100). However, there is insufficient evidence to identify a statistically significant difference (unequal variance *t* test on ranked data, *P*=.65). As with the observed mean number of included studies as a percentage of search results, the authors urge caution in concluding that no difference exists, and discuss possible reasons in the following section.

Discussion

Principal Findings

As observed from the reported *P* values in Table 7, there is very strong evidence that the number of articles at each stage of the screening process is higher for DTA reviews than for those of treatment, in support of hypothesis A (and hypothesis C in the case of an increased number of raw search results). This demonstrates a significant increase in the required workload for systematic reviews of diagnostic test accuracy. In addition, very strong evidence is obtained in support of hypothesis B. However, the p-values obtained for both the number of included and full-text articles retrieved as a percentage of the total search results were insufficient to ascertain a statistically significant difference between the means for DTA and treatment reviews.

As reported in Table 5 and 6, the standard deviation for all results is quite large. In addition, our analysis is limited in that only 13 completed Cochrane DTA reviews existed as on the search date. This small sample size, combined with the large standard deviations results in relatively low power. There is a possibility that the negative results reported for the included and full-text articles as a percentage of total search results were type II errors. This possibility is enhanced by the relatively large magnitude of the differences in sample means (see Table 5). Of course, it is impossible to say for certain until more data is available.

The authors note that while the analysis does not support the claim of sub-optimal metadata for DTA reviews, such a claim is not new and is supported by previously published literature. In addition to the lack of a definitive MeSH term for DTA studies, the Cochrane Handbook for reviews of DTA studies [15] notes that many index and reference tests employed during DTA studies have no corresponding MeSH term. From the handbook: a “database of names used to describe index tests and reference standards is being built”. However it is not complete as yet and due to the size of databases like MEDLINE and EMBASE, it is unlikely to be able to be applied retrospectively.

The reported results (summarized in Table 7) combined with the substantial difference in observed effect sizes lead the authors to conclude that the analysis supports the claim that DTA reviews present additional IR challenges. The magnitude of the difference in effect sizes is of particular importance as it implies a practical difference in the level of effort required for DTA and treatment reviews. They note the limitations of the study due to the small sample size of available DTA reviews. Further analysis needs to be done when more data is available.

It is interesting to note that the expected manifestations of hypothesis B (increased target class heterogeneity) could be said to drive the expected increase in workload during stage 2 screening described in hypothesis A. Similarly, hypothesis C (sub-optimal metadata) could be said to drive the increased workload in stage 1. This provides an interesting guide to any future work on the application of data mining to DTA reviews; by addressing these challenges the comparative difficulty of DTA reviews can be reduced.

We also like to mention that the hypotheses discussed in this paper could have additional manifestations throughout the review in addition to those in the literature search and screening stages. For example, the increased range of study designs and analysis methodologies (hypothesis B) could lead to increased difficulty in performing or interpreting any subsequent meta-analysis. As the focus of this paper is the literature search/screening stages of DTA reviews (and due to the inability to observe such manifestations in our data) we do not consider

such manifestations in our work, however such a study in future may be interesting.

Strengths and Limitations

To the best of our knowledge, this is the first study directly comparing reviewer workload for literature screening between systematic reviews of DTA and treatment. In addition, as stated in the data collection section, basing the comparison of DTA and treatment samples off the reported number of citations screened (rather than rerunning searches where applicable) is an advantage of our study. Such an approach implicitly takes into account decisions by authors about the required sensitivity of the initial search, which can be expected to differ across individual reviews and clinical domains.

There are several limitations to our study. Firstly, the relatively small number of Cochrane DTA reviews published as of the search date (n=13) results in statistical analysis with low power. As more data is available, future studies that permit comparison of DTA and treatment reviews in fields beyond those published to date by the 10 CRGs, would be of interest. Our results may also be biased towards Cochrane Reviews, as our analysis was performed purely on reviews collected from the Cochrane database of systematic reviews. As discussed earlier in the article, we believe this decision to be justified as it helps restrict the variance between two samples to clinical type (ie, DTA or treatment). Nonetheless this needs to be considered when interpreting the results of our study.

Conclusions

We demonstrate an increase in practical difficulty when screening literature for DTA reviews as compared to treatment. In addition, some potential causes for this additional difficulty at each stage of the literature search process are presented. We make three main conclusions in this article: first, the overall reviewer workload during literature screening is higher for DTA reviews than for treatment, as evidenced by the larger number of citations obtained at each stage of the literature screening process. Second, the target class of studies included in DTA reviews is broader than the corresponding class for reviews of treatment, as evidenced by the lower number of included studies as a percentage of full-text articles screened. Finally, we provide partial statistical evidence to support claims of the relative unsuitability of available metadata for DTA reviews. We note that future analyses with higher statistical power would be of greater interest.

This article provides a strong case for increased attention from the IR community on systematic reviews of DTA. Such work to address the challenges discussed in this paper could lead to genuine reductions in the workload and difficulty of conducting DTA reviews. One possible direction for future research includes developing high quality classifiers for DTA studies. This could help build consensus with the goal of widespread use of methodological search filters, similar to the current practice for Cochrane Reviews of treatment. As authors for DTA reviews must take into account that relevant data for any meta-analysis can often be synthesized from a range of studies (for example, non-DTA studies reporting individualized patient data [32-35]), this task could be further refined to develop classifiers for things such as individual study designs (ie, cohort study, case-control study), or to simply identify studies that report things like individualized patient data for a given test. Another advantage of individual patient data is that it will allow for a more tailored application to clinical scenarios via subgroup analysis.

In addition, given the size and scope of resources such as MeSH, it is unreasonable to expect all relevant metadata to be assigned to all references. The development of classifiers to assign interesting or relevant MeSH terms would help to increase the recall of interesting terms, potentially allowing for the creation of shorter, more specific queries. Such classifiers could also be used to apply newer MeSH terms retrospectively in existing databases. Finally, a third potential direction includes the application of data mining to identify which MeSH terms have particularly high discriminative power for DTA reviews. This task works in conjunction with the development of MeSH classifiers. Alternatively, data mining could be applied to identify clusters of citations that do not correspond to specific MeSH terms but nonetheless contain good discriminative power.

Over time, as the above concerns are addressed it could be expected that the required workload for DTA and treatment reviews converge. However there are two reasons for which research addressing IR for reviews with very high levels of class-imbalance (such as those currently observed for DTA reviews) is also required: first, the number of references screened for systematic reviews is heavily right-tailed (see data collection). For both treatment and DTA, dealing with reviews at the extreme end of the spectrum is an open problem [9]. And second, while it can be expected that future developments in mitigating the above challenges will reduce the levels of class-imbalance, it is unlikely that an optimal solution will be found in the near future. In addition, while efforts are occasionally made to retrospectively update metadata for databases such as MEDLINE and EMBASE where a sufficient need can be demonstrated (eg, the MeSH re-tagging project for randomized controlled trials [36]), the cost and difficulty of such tasks implies that some challenges are unlikely to be entirely solved.

Acknowledgments

The authors would like to gratefully acknowledge Xin Zhang from the School of Medicine at the University of Sydney for her contributions to this project. We would also like to acknowledge the anonymous reviewers for numerous insightful comments and suggestions.

Conflicts of Interest

None declared.

Multimedia Appendix 1

List of Cochrane Reviews included in the analysis.

[XLS File (Microsoft Excel File), 68KB - medinform_v2i1e11_app1.xls]

References

1. Wright RW, Brand RA, Dunn W, Spindler KP. How to write a systematic review. Clin Orthop Relat Res 2007 Feb;455:23-29. [doi: 10.1097/BLO.0b013e31802c9098] [Medline: 17279036]
2. McLellan F. 1966 and all that-when is a literature search done? Lancet 2001 Aug 25;358(9282):646. [doi: 10.1016/S0140-6736(01)05826-3] [Medline: 11530164]
3. Sampson M, Shojania KG, Garritty C, Horsley T, Ocampo M, Moher D. Systematic reviews can be produced and published faster. J Clin Epidemiol 2008 Jun;61(6):531-536. [doi: 10.1016/j.jclinepi.2008.02.004] [Medline: 18471656]
4. Tsafnat G, Dunn A, Glasziou P, Coiera E. The automation of systematic reviews. BMJ 2013;346:f139. [Medline: 23305843]
5. Frunza O, Inkpen D, Matwin S. Building systematic reviews using automatic text classification techniques. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Beijing, China: Association for Computational Linguistics; 2010:303-311.
6. Cohen AM, Ambert K, McDonagh M. Cross-topic learning for work prioritization in systematic review creation and update. J Am Med Inform Assoc 2009;16(5):690-704 [FREE Full text] [doi: 10.1197/jamia.M3162] [Medline: 19567792]
7. Wallace BC, Small K, Brodley CE, Trikalinos TA. Active learning for biomedical citation screening. 2010 Presented at: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discoverydata mining; July 25-28, 2010; Washington, DC, USA p. 173-182. [doi: 10.1145/1835804.1835829]
8. Cohen AM. Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@95 measure. J Am Med Inform Assoc 2011;18(1):104-105 [FREE Full text] [doi: 10.1136/jamia.2010.008177] [Medline: 21169622]
9. Matwin S, Kouznetsov A, Inkpen D, Frunza O, O'Blenis P. Performance of SVM and Bayesian classifiers on the systematic review classification task. J Am Med Inform Assoc 2010 Dec 18;18(1):104-105. [doi: 10.1136/jamia.2010.009555]
10. Devillé WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2002 Jul 3;2:9 [FREE Full text] [Medline: 12097142]
11. Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews. Res Syn Meth 2011 Apr 11;2(1):1-14. [doi: 10.1002/jrsm.27]
12. Ananiadou S, Rea B, Okazaki N, Procter R, Thomas J. Supporting Systematic Reviews Using Text Mining. Soc Sci Comput Rev 2009 Apr 20;27(4):509-523. [doi: 10.1177/0894439309332293]
13. Pai M, McCulloch M, Enanoria W, Colford JM. Systematic reviews of diagnostic test evaluations: What's behind the scenes? ACP J Club 2004;141(1):A11-A13. [Medline: 15230574]
14. Higgins JPT, Green S. Cochrane Collaboration. In: Cochrane handbook for systematic reviews of interventions. Chichester, England: Wiley-Blackwell; 2008.
15. Deeks JJ, Wisniewski S, Davenport C. Guide to the contents of a Cochrane Diagnostic Test Accuracy Protocol. In: Deeks J, editor. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. London: The Cochrane Collaboration; 2013.
16. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM, Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008 Dec 16;149(12):889-897 [FREE Full text] [Medline: 19075208]
17. Kastner M, Wilczynski NL, McKibbon AK, Garg AX, Haynes RB. Diagnostic test systematic reviews: bibliographic search filters ("Clinical Queries") for diagnostic accuracy studies perform well. J Clin Epidemiol 2009 Sep;62(9):974-981 [FREE Full text] [doi: 10.1016/j.jclinepi.2008.11.006] [Medline: 19230607]
18. Leeflang MM, Scholten RJ, Rutjes AW, Reitsma JB, Bossuyt PM. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. J Clin Epidemiol 2006 Mar;59(3):234-240. [doi: 10.1016/j.jclinepi.2005.07.014] [Medline: 16488353]
19. Whiting P, Westwood M, Beynon R, Burke M, Sterne JA, Glanville J. Inclusion of methodological filters in searches for diagnostic test accuracy studies misses relevant studies. J Clin Epidemiol 2011 Jun;64(6):602-607. [doi: 10.1016/j.jclinepi.2010.07.006] [Medline: 21075596]
20. Kastner M, Haynes RB, Wilczynski NL. Inclusion of methodological filters in searches for diagnostic test accuracy studies misses relevant studies. J Clin Epidemiol 2012 Jan;65(1):116-117. [doi: 10.1016/j.jclinepi.2011.02.011] [Medline: 22118266]
21. Wilczynski NL, McKibbon KA, Walter SD, Garg AX, Haynes RB. MEDLINE clinical queries are robust when searching in recent publishing years. J Am Med Inform Assoc 2013;20(2):363-368 [FREE Full text] [doi: 10.1136/amiajnl-2012-001075] [Medline: 23019242]
22. Ritchie G, Glanville J, Lefebvre C. Do published search filters to identify diagnostic test accuracy studies perform adequately? Health Info Libr J 2007 Sep;24(3):188-192. [doi: 10.1111/j.1471-1842.2007.00735.x] [Medline: 17714173]

23. Karimi S, Zobel J, Pohl S, Scholer F. The challenge of high recall in biomedical systematic search. : ACM; 2009 Presented at: Proceedings of the Third International Workshop on Datatext Mining in Bioinformatics; November 2-6, 2009; Hong Kong, China p. 89-92. [doi: 10.1145/1651318.1651338]
24. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc 2006;13(2):206-219 [FREE Full text] [doi: 10.1197/jamia.M1929] [Medline: 16357352]
25. Matwin S, Kouznetsov A, Inkpen D, Frunza O, O'Blenis P. A new algorithm for reducing the workload of experts in performing systematic reviews. J Am Med Inform Assoc 2010;17(4):446-453 [FREE Full text] [doi: 10.1136/jamia.2010.004325] [Medline: 20595313]
26. Cohen AM. Optimizing feature representation for automated systematic review work prioritization. AMIA Annu Symp Proc 2008;2008:121-125 [FREE Full text] [Medline: 18998798]
27. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics 2010;11:55 [FREE Full text] [doi: 10.1186/1471-2105-11-55] [Medline: 20102628]
28. Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. : ACM; 2012 Presented at: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium; January 28-30, 2012; Miami, Florida, USA p. 819-824. [doi: 10.1145/2110363.2110464]
29. Cohen AM, Ambert K, McDonagh M. Studying the potential impact of automated document classification on scheduling a systematic review update. BMC Med Inform Decis Mak 2012;12:33 [FREE Full text] [doi: 10.1186/1472-6947-12-33] [Medline: 22515596]
30. Trieschnigg D, Pezik P, Lee V, de Jong F, Kraaij W, Rebholz-Schuhmann D. MeSH Up: effective MeSH text classification for improved document retrieval. Bioinformatics 2009 Jun 1;25(11):1412-1418 [FREE Full text] [doi: 10.1093/bioinformatics/btp249] [Medline: 19376821]
31. Aphinyanaphongs Y, Aliferis CF. Text categorization models for retrieval of high quality articles in internal medicine. AMIA Annu Symp Proc 2003;2003:31-35 [FREE Full text] [Medline: 14728128]
32. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health Technol Assess 2005 Mar;9(12):1-113, iii [FREE Full text] [Medline: 15774235]
33. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med 2002 Jun 15;21(11):1525-1537. [doi: 10.1002/sim.1185] [Medline: 12111918]
34. Riley RD, Dodd SR, Craig JV, Thompson JR, Williamson PR. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. Stat Med 2008 Dec 20;27(29):6111-6136. [doi: 10.1002/sim.3441] [Medline: 18816508]
35. Ter Riet G, Bachmann LM, Kessels AG, Khan KS. Individual patient data meta-analysis of diagnostic studies: opportunities and challenges. Evid Based Med 2013 Oct;18(5):165-169. [doi: 10.1136/eb-2012-101145] [Medline: 23704701]
36. Glanville JM, Lefebvre C, Miles JN, Camosso-Stefinovic J. How to identify randomized controlled trials in MEDLINE: ten years on. J Med Libr Assoc 2006 Apr;94(2):130-136 [FREE Full text] [Medline: 16636704]

Edited by G Eysenbach; submitted 21.10.13; peer-reviewed by T de Bray, S Kim, B Skidmore; comments to author 10.11.13; revised version received 12.03.14; accepted 22.04.14; published 27.05.14.

Please cite as:
Petersen H, Poon J, Poon SK, Loy C
Increased Workload for Systematic Review Literature Searches of Diagnostic Tests Compared With Treatments: Challenges and Opportunities
JMIR Med Inform 2014;2(1):e11
URL: http://medinform.jmir.org/2014/1/e11/
doi:10.2196/medinform.3037
PMID:25600450

Original Paper

Electronic Clinical Safety Reporting System: A Benefits Evaluation

Pamela Elliott[1], BN, MBA, PhD; Desmond Martin[2], BA, BSc, MSc; Doreen Neville[3], BN, MScN, SM, ScD

[1]Elliott's Enterprises, St John's, NL, Canada

[2]Eastern Health, Department of Research, St John's, NL, Canada

[3]Memorial University, Office of the Provost and Vice-President (Academic), St John's, NL, Canada

Corresponding Author:
Pamela Elliott, BN, MBA, PhD
Elliott's Enterprises
320 Torbay Road, Suite 204
St John's, NL, A1A 4E1
Canada
Phone: 1 709 754 6706
Fax: 1 709 579 2652
Email: pelliott@nl.rogers.com

Abstract

Background: Eastern Health, a large health care organization in Newfoundland and Labrador (NL), started a staged implementation of an electronic occurrence reporting system (used interchangeably with “clinical safety reporting system”) in 2008, completing Phase One in 2009. The electronic clinical safety reporting system (CSRS) was designed to replace a paper-based system. The CSRS involves reporting on occurrences such as falls, safety/security issues, medication errors, treatment and procedural mishaps, medical equipment malfunctions, and close calls. The electronic system was purchased from a vendor in the United Kingdom that had implemented the system in the United Kingdom and other places, such as British Columbia. The main objective of the new system was to improve the reporting process with the goal of improving clinical safety. The project was funded jointly by Eastern Health and Canada Health Infoway.

Objective: The objectives of the evaluation were to: (1) assess the CSRS on achieving its stated objectives (particularly, the benefits realized and lessons learned), and (2) identify contributions, if any, that can be made to the emerging field of electronic clinical safety reporting.

Methods: The evaluation involved mixed methods, including extensive stakeholder participation, pre/post comparative study design, and triangulation of data where possible. The data were collected from several sources, such as project documentation, occurrence reporting records, stakeholder workshops, surveys, focus groups, and key informant interviews.

Results: The findings provided evidence that frontline staff and managers support the CSRS, identifying both benefits and areas for improvement. Many benefits were realized, such as increases in the number of occurrences reported, in occurrences reported within 48 hours, in occurrences reported by staff other than registered nurses, in close calls reported, and improved timelines for notification. There was also user satisfaction with the tool regarding ease of use, accessibility, and consistency. The implementation process encountered challenges related to customizing the software and the development of the classification system for coding occurrences. This impacted on the ability of the managers to close-out files in a timely fashion. The issues that were identified, and suggestions for improvements to the form itself, were shared with the Project Team as soon as they were noted. Changes were made to the system before the rollout.

Conclusions: There were many benefits realized from the new system that can contribute to improved clinical safety. The participants preferred the electronic system over the paper-based system. The lessons learned during the implementation process resulted in recommendations that informed the rollout of the system in Eastern Health, and in other health care organizations in the province of Newfoundland and Labrador. This study also informed the evaluation of other health organizations in the province, which was completed in 2013.

(JMIR Med Inform 2014;2(1):e12) doi:10.2196/medinform.3316

KEYWORDS

electronic occurrence reporting; electronic clinical safety reporting; adverse event reporting in health care; evaluating electronic reporting systems in health care; health information technology evaluations

Introduction

The Risks of Health Care

Florence Nightingale once wrote, "it may seem a strange principle to enunciate as the very first requirement in a hospital that it should do the sick no harm" [1]. That was over a hundred and fifty years ago, and yet, today that requirement of "doing no harm" is still identified as an issue in the health system. While the health system has changed since that time, the "doing no harm" to patients is part of the patient safety agenda worldwide in health care.

Health care is provided in a high risk environment. In a report by The National Steering Committee on Patient Safety [2], which outlines a strategy for improving patient safety in Canadian health care, a brief description of that high risk environment is provided,

> *Health care is provided 24 hours a day, seven days a week. Dramatic advances in the diagnosis and treatment of disease have made care processes more complex; however, many organizations are hampered by outdated modes of communication, record keeping, employee training, and traditional hierarchical authority structures. The aging population, resource limitations, a critical shortage of qualified health professionals in a growing list of locations and specialties, and challenges created by mergers, and restructuring within health care organizations, are creating unequalled strain on the systems, thus, increasing the likelihood of adverse events, sometimes with lethal consequences.* [The National Steering Committee on Patient Safety, [2], p 5]

Patient safety has been defined in the Canadian Patient Safety Dictionary as "the reduction and mitigation of unsafe acts within the health care system, as well as through the use of best practices shown to lead to optimal patient outcomes" [3].

Patient Safety in Hospitals

The issue of patient safety has gained an increasing profile in recent years, especially since the publication of *To Err Is Human* by the Institute of Medicine in 2000. The report estimated that between 44,000 and 98,000 Americans die each year from adverse events at a cost to the nation of US $8.5 to $19 billion annually [4]. Other countries, including the United Kingdom, Australia, and New Zealand have investigated the extent of the problem, and clearly shown that adverse events are a global patient safety concern [5-9]. Baker et al [5] conducted a detailed study of patient safety in Canada, and revealed that 7.5% of adult acute care patients in Canadian hospitals in the year 2000 experienced an adverse event, and 36.9% of these events were deemed to be preventable. The study estimated that between 9250 and 13,750 deaths from adverse events could have been prevented. Their study also looked at similar studies in other countries (United Kingdom, Australia, New Zealand, and the United States), and found that adverse event rates ranged from 2.9% to 16.6% of acute care admissions. They point out that one of the key steps in promoting patient safety is to have a reporting system that allows adverse events and near misses/close calls to be recorded so that health care workers can learn from them and implement corrective action plans.

The development of reporting systems for adverse events in health care can be traced back to the late 1970s. Since then, many countries have been implementing reporting systems and moving to electronic systems. However, countries such as the United Kingdom, Australia, Japan, and the United States are ahead of other countries, including Canada, particularly as it relates to national reporting systems [7,10]

Eastern Health

Eastern Health is the largest integrated health organization in Atlantic Canada, serving a regional population of more than 290,000, and offering tertiary level and specialty services to a population of about 500,000 across the province of Newfoundland and Labrador. Eastern Health was formed in 2005 as a merger of seven organizations. The organization has approximately 12,000 staff members, and operates 27 institutional health service facilities and community health services in 30 communities. The services provided by Eastern Health cover a wide range of services in three sectors: (1) acute, (2) long term, and (3) community [11].

Clinical safety reporting is used interchangeably with occurrence reporting at Eastern Health, and refers to a process that facilitates the identification and monitoring of adverse events and incidents that occur during health care treatment or service and/or within health care facilities. The reporting system is used to report on occurrences such as falls, safety/security issues for patients, medication errors, treatment and procedural mishaps, and medical equipment malfunctions. This is consistent with the definition and approach outlined in the report of the provincial Task Force on Adverse Health Events [12]. An individual who is involved in an occurrence or witnesses an occurrence completes a report form and forwards it to the manager. The form captures information such as the patient name, patient record number, diagnosis, location of the incident, type of occurrence, time of occurrence, impact on patient, notification information, assessment information, physician assessment, and follow-up actions required. The form is only one part of the reporting system. The manager has the primary responsibility for ensuring the communication gets to the appropriate levels of authority and ensuring appropriate follow-up action. Depending on the complexity of the occurrence, and the follow-up actions required, the process can take from a few minutes to a few days, particularly if much consultation has to take place in determining the resolutions.

Early in the newly merged organization, Eastern Health recognized the need to improve and standardize it's occurrence reporting processes. Each of the organizations involved in the merger had their own reporting processes, most of which were paper based. There were issues with the paper reporting systems, such as inconsistencies in what was being reported, different forms in use throughout the region, delays in notification to the Quality and Risk Management (QRM) Department, incomplete forms, and lack of feedback to employees about what was being done to address the clinical safety issues identified [12]. In an effort to improve the reporting system, Eastern Health submitted a proposal to Canada Health Infoway, seeking funding to

implement an electronic occurrence reporting system. Canada Health Infoway is a national organization with the mandate for promoting the implementation of electronic records in the health system throughout the country. The proposal identified 13 objectives, all aimed at the ultimate goal of improving clinical safety.

Canada Health Infoway approved the funding to implement an electronic occurrence reporting system in late 2007, and the implementation began in 2008. The project required the selection of a vendor for the software applications for the Web-based system. The vendor chosen was based out of the United Kingdom, where many of its hospitals were using the system. Also, in Canada, the province of British Columbia (BC) had chosen the same vendor, and other health organizations were considering the same system. The software has the ability for organizations to customize some of their processes and terminology used in the occurrence reporting. The software chosen was anticipated to not only provide a user-friendly, confidential electronic form, but also help with other parts of the reporting process, such as the timely notification of the managers, improved communications between the different personnel involved, trending, and tracking.

Eastern Health is so large, that a staged implementation over several years was planned. The project budget included funding for a comprehensive evaluation of Phase One of the implementation, which involved four sites: (1) acute care, (2) long term care, (3) community health in an urban setting, and (4) an integrated services site in a rural setting. The evaluation was designed with the goal of determining if the anticipated benefits were realized, and if there were any lessons learned that could help with future implementations. The evaluation study was completed in 2010. The full evaluation was much more comprehensive in scope than presented in this paper. This paper outlines the evaluation approach used, and focuses on the key findings, particularly the benefits realized.

Methods

Evaluation of Health Information Technology

The evaluation of information technology (IT) in heath care is not conducive to the methods used in laboratory systems or “gold standards”, such as randomized controlled methods. Therefore, being able to identify causality is a limitation. A particular limitation is the difficulty in measuring or controlling for confounding variables, variables that are associated with the exposure of interest and also with the outcome of interest [13]. The physical settings, type of clinical service, acuity of patients, practices, and the experience of providers is not conducive to randomization and setting up control groups. Also, an important part of the evaluation of electronic health information systems is the end users’ acceptance of the system, and lessons learned which could assist in other implementations or system enhancements. Multi-methods, including the pre- and post-testing of interventions, is often advocated in health care IT evaluations. This quasi-experimental design is often used in the evaluation of health information systems due to time, cost, and technical restraints.

Approach to This Evaluation

The approach to this evaluation was extensive, using both qualitative and quantitative methods. The design in this study involved measuring occurrence reporting data for a 6 month period before the implementation, and six months post implementation, as well as pre- and post-qualitative data. The design also involved a post test regarding user satisfaction, as well as the evaluation of training sessions.

The approach, including the development of data collection tools, was informed by five previous works in the evaluation of electronic health information systems and in patient safety. First, the work of Neville et al [14], which outlines a framework for evaluating electronic health records initiatives. The framework involves stakeholders throughout the process and utilizes a pre- and post-study design. Second, the work of Delone and McLean [15] on an information system success model, which has been incorporated by Canada Health Infoway into a benefits evaluation framework by Lau et al [16]. A key component of this framework involves the identification of indicators that can be used in the development of data collection tools to measure various dimensions of information systems success. Third, the work conducted by the British Columbia Electronic Incident Reporting Pilot Project in evaluating the same reporting system implemented at Eastern Health [17]. Fourth, the work of Ginsburg et al [18] and Accreditation Canada [19] in patient safety culture, and finally, pre-evaluation workshops attended by key stakeholders, which focused on the identification of research questions and indicators of interest.

The full evaluation study for the project focused on the following research questions, and used the data sources as outlined in Table 1. The scope of this paper is reporting on just a part of the larger evaluation, mainly on the benefits realized and the disadvantages/areas for improvement.

Table 1. Research questions and data sources used.

Research questions	Data sources
1) Anticipated benefits of this system.	Stakeholder workshops Project documents Literature review Focus groups Key informant interview
2) Benefits achieved and comparison with anticipated benefits.	Surveys Administrative records Focus groups Key informant interviews
3) Projected costs of this system.	Project documents
4) Costs of implementing the system and comparison with projected costs.	Project documents and discussion with implementation team
5) Necessary planning and management structures in place to proceed with the project.	Key informant interviews Focus groups Discussion with implementation team
6) Unforeseen harms and/or disadvantages.	Key informant interviews Focus groups
7) Key facilitators and barriers to successful implementation of the project.	Key informant interviews Focus groups Surveys Project documents

The Indicators

The indicators chosen were based on the feedback that was obtained at a stakeholder workshop, and a review of the literature. Even though the full evaluation focused on many indicators, this paper will highlight the findings for the key indicators as follows: (1) number of occurrences reported, (2) reporter characteristics (nurses and non nurses), (3) timelines for reporting, (4) user satisfaction, (5) perceived benefits, and (6) perceived disadvantages.

The occurrence reporting data was compared 6 months post implementation to a similar 6 months period pre-implementation for each of the sites.

All of the frontline clinical staff and managers working in each of the four sites were included in the sampling for the user satisfaction questionnaires. These included staff such as registered nurses (RN), licensed practical nurses, personal care attendants, allied health professionals, ward clerks, diagnostic imaging, and laboratory staff. The physicians, research, and nondirect care staff were excluded from the sample. The rationale for the inclusions and exclusions was based on the historical utilization of occurrence reporting, and the planned implementation schedule. The user satisfaction survey questionnaire had close-ended questions, mainly about the electronic tool, and used a Likert-type scale.

The sampling for the interviews included all of the senior managers involved with the new system. The sampling for the focus groups included all of the frontline managers and frontline clinical staff at the sites who were using the new system. The pre- and post-focus groups and key informant guides used open-ended questions, focusing mostly on the perceived benefits and disadvantages/suggestions for improvement, as well as the facilitators and barriers.

Results

Response

Participation was voluntary for taking the satisfaction survey. There were 1074 user satisfaction surveys distributed post implementation, with 358 staff (330 frontline staff and 28 managers) responding for a response rate of 33.33%. Of the 358 who responded, 205 (57.3%) had used the system. The questionnaires were sent to the same staff pre- and post-implementation. Due to the nature of occurrence reporting, not all staff would have been involved in using the system during the study period, unless they had experienced or witnessed an occurrence. It is the staff that used the system that provided the data for the analysis related to the user satisfaction of the tool itself.

There were pre- and post-key informant interviews conducted, with 11 senior managers participating in both. There were pre- and post-focus groups conducted with the frontline managers and staff, with 12 managers and 13 frontline staff participating in the post implementation focus groups, as well as focused discussions with the project team. The qualitative results of all these groups and interviews contributed to the data discussed in this paper. A limitation is that there was low participation of frontline staff in the focus groups, even though the focus groups were held at lunchtime with lunch provided. Posters and email notices were provided at each site, but there was little response.

Findings

In addition to the user satisfaction surveys, focus groups, and interviews, there was also a review of occurrence reporting administrative records for a 6 month period pre-implementation and 6 month post implementation. See Table 2, which compares the occurrence reporting data for a 6 month period before the implementation, to a 6 month period following the implementation.

Table 2. Comparison of the 6 months pre-implementation and the 6 months post implementation data.

Occurrence reports indicators	Pre-implementation (%)	Post-implementation (%)	Change/improvement between pre- and post-implementation (%)
# of occurrences reported	495	907	Increase 412 reports (83.2)
Reports completed	386 (77.9)	795 (87.6)	Increase (9.7)
Non-RN reports	129 (26.1)	391 (43.1)	Increase (17.0)
Reported within 48 hours of occurrence	166 (33.6)	799 (88.1)	Increase (54.5)
Average time between occurrence and notification of the manager sign-off	11.3 days	17 days	Increase 5.7 days (50.4)
Average time between occurrence and notification of quality and risk management	43 days	Immediately	Decrease 43 days (100.0)
Close calls	5 (1.0)	97 (10.7)	Increase (9.7)

Key Benefits

The main findings of this study show that there are several key benefits realized, such as increased reporting of occurrences, improvement in the number of reports completed, more reporting by non-RN health care employees, improved notification of the managers and the QRM Department, and increased reporting of close calls. There were also some challenges experienced, such as decreased time in some areas for the close-out/sign-off of files. In addition to the changes in reporting, there was also satisfaction expressed by users with the new system.

The results of the user satisfaction surveys show that respondents across all care settings seem to be satisfied with the new electronic system. They report that the system is easy to use and consistent in performance. Other benefits identified in focus groups and interviews included: (1) easy access to computers and forms, (2) improved legibility, (3) increased awareness of what constitutes an occurrence and close call, (4) less time required to complete reports, (5) availability of information about the status of the individual managers' occurrences, (6) easy to complete forms, (7) less paper shuffling, (8) more detailed information on reports, (9) easier to track follow-up actions, (10) improved confidentiality (reports not lying around at a nursing station for others to see), and (11) fewer misplaced reports. While all occurrence reports (paper or electronic) are expected to be confidential, paper reports are more vulnerable to being viewed by more than those involved in the occurrence. The electronic tool requires a password for access, and only those involved in completing the report, follow-up actions, and/or quality risk management personnel are permitted to view them.

Areas for Improvement

Even though there were mostly positive comments about the reporting form, and most employees said they liked it, several areas for improvement were mentioned by the frontline staff. These included: (1) no place on the form for the person who attends to the client, the intervention, or a physician section to make notes; (2) form is too long; (3) locator drop down box does not lend itself to identifying the exact location of the occurrence (for example, "the room number"); and (4) the "locator function takes too long to scroll down to find the area of the occurrence". Some participants also mentioned that sometimes there is not much feedback on the form from their managers regarding the follow-up action and prevention measures taken, however, they did indicate that they now have a reference file number for the report, and can see that it was reviewed.

Discussion

User Satisfaction

Many of the benefits identified are consistent with those identified in other studies. While the participants were not asked to prioritize or rank benefits, the ease of use was the most commonly mentioned. This is a similar finding to other studies with ease of use being the most frequently cited benefit [20-28]. Other benefits, such as those found in this study, are less cited. This study also included benefits not identified in the literature reviewed, such as the availability of information about the status of the individual manager's occurrences, and fewer misplaced occurrence reports.

Even though many benefits were identified, there were several points of dissatisfaction raised by end users. For the management group, the inability to close-out files, and the uncertainty about whether or not the file was closed, were viewed as undesirable. When a report was changing handlers (a term used to describe the person following up on the report), they were unsure as to what happened with the report, as there was no confirmation if the handler received or acted on the report. There was also confusion at times with respect to management responsibility

for a particular report when an occurrence involved two departments and one employee. This inability to "close the gap" was a concern, because the managers felt that despite the fact that they had taken appropriate follow-up action, it was not showing in the system in a timely fashion. There was also recognition that the implementation was not yet completed with respect to the coding classification of occurrences, and consequently, the managers were not able to get timely customized reports. At the time of this evaluation, work on these issues by the Project Implementation Team was in progress, and the managers indicated that addressing these functionality issues and getting the reports would enhance their view of the system. Although these issues were raised in the focus groups and interviews, the managers who responded to the user satisfaction survey expressed a high degree of satisfaction with the tool.

The issue of the locator function on the form itself was similar to a finding from a study on the same system by Walsh and Antony [29], where the location of the incidents was identified as a concern. The locator function is a feature that can be customized to the setting. The issue identified in the interviews and focus groups with the lack of customized "drop down boxes" was for specialized areas, such as laboratory and pharmacy services. The staff from the nursing areas, however, indicated satisfaction with the drop down boxes. The Project Implementation Team reported that there is a plan to customize the drop down boxes for the clinical support areas (eg, the Diagnostic Imaging, Laboratory, and Pharmacy Departments) to assist in making them user-friendlier for all end users.

As noted in the Findings section, the participants reported that there is no place on the form for employees to receive the feedback from their managers regarding the detailed follow-up actions and prevention measures taken. Other studies [24-26,30,31] point to the importance of feedback to staff, and that staff want to see that by taking the time to report an occurrence, there will be corrective action taken, and that quality will improve. It is well recognized that "you cannot fix what you cannot measure". However, Clarke et al [30] point out that it is important to be aware of the types of problems that need to be fixed, rather than focus on all the instances of problems that need to be counted (p. 314). The counting can be used in tracking, but must be accompanied by action. The importance of receiving feedback on occurrences, and ensuring that corrective action is taken, was a common theme for both the managers and the frontline staff in this study.

Changes in Reporting

There were notable increases in the numbers of occurrences reported in all settings, which is consistent with the findings from other studies [20,22,26,32]. While the number of occurrences increased across all sectors, it is difficult to analyze data about the types of occurrences across sectors. A review paper by Boxwala et al [33] examined various approaches to identifying errors and adverse events (of which occurrence reporting is one), and cautions about making any comparisons across sectors on the numbers and types of incidents, as there are factors such as inconsistent patient safety terminology, the clinical context including the roles of various personnel in the incident, the location, and other contributing factors.

A detailed breakdown of the types of occurrences reported by providers was not conducted. However, a high level review showed that there was a large increase (from 5 to 160) in the number of occurrences reported in the Clinical Assessment category. This category includes incomplete information on a requisition and/or specimen. This is consistent with the increase in reporting by the Diagnostic Services staff (radiology and laboratory), and was also mentioned in the focus groups and interviews. As in the pre-implementation period, the nurses were still the highest reporters for the Falls and Medications categories.

In a study by Zboril-Benson and Magee [34], there was an improvement in the types of incidents reported in a pilot project after cultural and educational changes were made. Pre-pilot reports at their study site indicated that only serious errors in health care were likely to be reported (ie, when a patient has been injured; when a willful violation of established protocol has been violated, etc). After the delivery of education sessions, they found an increase in the reporting of both close calls and occurrences with no harm.

The findings in the Zboril-Benson and Magee study are similar to the findings in this study. In the focus groups and key informant interviews, the participants indicated the education sessions that were conducted as part of the implementation process contributed to a better understanding and heightened awareness of the importance of reporting all occurrences and close calls. There was an increased awareness of what constitutes an occurrence. The participants indicated a better awareness of how the reporting of close calls can lead to system improvements.

An explanation given by a manager in this study, for an increase in reporting, was that even though all staff members in the paper-based system were expected to report occurrences (even when there was no harm to the patient), they did not, and often they just dealt with the issue. An example provided was that of a missing armband, "the staff would just do another armband for the patient and not write up the report". The participants reported an increased understanding of how the tracking and trending of occurrences (even when there is no harm) can contribute to policy and practice changes.

Another contributing factor to the increase in reporting is the improved satisfaction expressed by employees with the ease of use and accessibility of the electronic tool. As was stated in the focus groups, "If staff members are busy, they may not bother to take the time from their day to find a paper report form and write up the occurrence, especially if no harm resulted to the patient". The fact that the new system also provided a feedback mechanism to the reporter was identified as a benefit. Many reported that in the past, they completed reports, but often never heard back about what was done with the report or about the issues identified. As one participant described "it is like the report went in to the big black hole". Now, it is easier to check on the status of the report, as they are given a file number when they log on and complete the electronic form.

Reporter Characteristics

This study found a notable change in reporter characteristics post implementation of the electronic system, moving beyond the traditional RN reporter, from 129/495 (26.1%) to 391/907 (43.1%) of occurrences reported by other than RNs. Even though RNs were still the main reporting health care worker group, other workers, such as allied health professionals, ward clerks, medical records, diagnostic, and laboratory staff also reported more occurrences.

Nurses are still the most frequent reporters, and the most frequent types of occurrences are related to falls, medication administration, and safety/security issues in the clinical settings. This finding is consistent with those of previous research [22,26,35-37]. Blais et al [36] point out that because "nurses are often the professionals who fill out the incident report forms, the adverse events they report on are generally limited to the problems relevant to their work" (p. 11). Most other studies reviewed for this study focused on acute care. In all settings in this study (acute care, community, and long term care), nurses were still the predominant reporting category.

Timelines for Reporting

There were improvements related to the timing of the notification of the occurrence to the QRM Department, and to the various management levels. In the past, the more serious occurrences, which resulted in harm to the patient, were usually reported as soon as possible, but often with occurrences that were of a lesser consequence, the reports were just sent over in the mail or notifications done when the manager could "get around to it". The tool is designed to produce the immediate notification of the occurrence to the manager and the QRM Department, and can be customized for notification alerts to different managers, depending on the needs of the area. This immediate notification function was identified by the managers as one of the key benefits of the electronic system, as it improves the efficiency of the communication channels in the organization with respect to notification about occurrences. This finding is consistent with the Cochrane et al [22] study. The improved notification features also contributed to the increased number of occurrences reported within 48 hours of the occurrence. The increase in this study was 54.5% (88.1% -33.6% from Table 2 above) compared to the Cochrane study, which was 82%, the difference in the magnitude, being related to the difference in pre-implementation baseline timelines, where the Cochrane study was much lower on this indicator. The post implementation timeline was similar for both studies, with 799/907 (88.1%) being the result in this study, and 84% being the result in the Cochrane study.

Post implementation, there was an increase in the average time (5.7 days) for the managers to sign-off on the report, compared to the previous paper-based reporting system, going from 11.3 to 17 days. The managers, quality leaders, and project leadership indicated that the decreased efficiency was related to the increase in the number of occurrences reported, as well as to the learning curve of the managers using the system. This new system resulted in an increased demand for follow-up activity, especially in areas where the number of occurrences had increased significantly, mainly in acute care, and the managers reported getting behind in completing files due to the other many competing demands on their time. The managers reported difficulties in understanding how to sign-off on the occurrence (follow-up completion), and they were not sure if they were completing this function correctly. As a result, the occurrence reports follow-up process, and closing-out the file were longer to complete overall. Hence, the system did not improve efficiency on this indicator during the 6 month post implementation period. This is in contrast to the study by Cochrane et al [22], where the average time between the event and the completion of the investigation decreased by 6 days, going from 39 days to 33 days. The baseline data was different. The time required for completing the investigation in BC (33 days) is longer than closing out a file in this study (17 days), but comparisons are difficult as the policies and procedures for closing out versus completing an investigation may differ, as well as the types of occurrences reported. The researchers in the Cochrane study [22] felt their result to be "only a slight improvement due to two factors: (1) the setting where the study took place was a busy unit where the manager had to support clinical work with limited opportunity to perform nonclinical, nonurgent work, which included doing follow-up work related to occurrence reports; and (2) the change in practice required of the manager was greater than anticipated" (p. 151). This was consistent with some of the feedback reported in this study. The managers reported that, in the past, they would "save up" the occurrence reports to complete them on "paper days", when they could have uninterrupted time. The new system provides immediate notification; however, obtaining uninterrupted time in a busy setting to focus on the follow-up actions is a challenge. This did impact on the sign-off/close-out time, and thereby is being perceived as a disadvantage with the new system.

Conclusions

This study showed that there are benefits to moving from a paper-based reporting of occurrences in health care to an electronic Web-based system. Some of the key benefits realized were an increase in the reporting of occurrences and close calls, improved timelines for notifying the managers and the QRM Department, improved tracking, more categories of the staff getting involved in reporting, reporting tool is easier to use, improved legibility, improved confidentiality, decreased reports missing, and more detailed information on the reports. It is important to point out that the implementation included extensive promotion and education of the new system, and this impacted on the awareness of employees, as identified in the interviews and focus groups. This, coupled with an easy to use electronic reporting tool, contributed to the benefits. The managers indicated that over time, the benefits realized would provide improved information that can lead to better tracking, trending, and addressing the clinical safety issues identified. While measuring the long term impact on clinical safety was beyond the scope of this evaluation, there was optimism expressed by the participants that if the employees continue to be engaged with the new system, then it will lead to improved clinical safety, as long as the issues identified are followed through with action plans. It would be interesting to repeat this study to see if the benefits realized after the first 6 months are sustained.

Most of the findings in this study are consistent with similar studies on Web-based electronic occurrence reporting systems in the acute care sector. The body of literature on the topic of benefits evaluations of electronic reporting systems in the acute care setting is small, and even smaller for the other sectors (community health and long term care). This study did include the long term care and community sectors, as well as the acute care settings, and the findings showed that there is little difference in the benefits realized between settings. The site was small for long term care (urban); therefore the findings from long term care settings have limitations. While some of the findings may be limited by the low participation of the frontline workers in the focus groups, the triangulation of the data from surveys, focus groups, interviews, and occurrence reporting records, provided evidence that there are benefits that can help in the pursuit of improved clinical safety, and that the employees support the system. The use of focus groups and key informant interviews provided information that was used to make improvements to the process and the tool.

The findings from this study were used to inform the rollout to the other sites at Eastern Health, and the implementation of the system in other health organizations in the province of Newfoundland and Labrador. There were changes made to the software and the implementation process based on the feedback obtained during the evaluation process. Also, the evaluation framework used in this study was used to guide the evaluation of the system in other regions in the province, which was completed in 2013. The evaluation approach for the provincial system used many of the same data collection tools as this study, but the amount of data collected was tailored to meet the resources available, as conducting pre- and post-studies can be quite costly. The evaluation tools and approach in this study have the potential to be used by other organizations that have the same or similar Web-based occurrence reporting systems.

Acknowledgments

The authors wish to acknowledge Eastern Health and Canada Health Infoway. Both organizations provided financial support for the project and the evaluation. The publication costs for this article have been covered by Canada Health Infoway Inc, a non-profit corporation funded by the Government of Canada. The funders had no role in study design, data collection and analysis, or decision to publish the manuscript. We also wish to acknowledge the many staff of Eastern Health who took time out of their busy schedules to participate in the surveys, focus groups, and interviews.

Conflicts of Interest

None declared.

References

1. Nightingale F. Florence Nightingale quotes. URL: http://womenshistory.about.com/cs/quotes/a/qu_nightingale.htm [accessed 2014-06-04] [WebCite Cache ID 6Q54U2zxo]
2. National Steering Committee on Patient Safety. HPC Network. Ottawa, ON: National Steering Committee on Patient Safety; 2002. Building a safer system: A national integrated strategy for improving patient safety in Canadian health care URL: http://hpclearinghouse.net/files/folders/other4/entry9427.aspx [accessed 2014-05-26] [WebCite Cache ID 6PusLedkl]
3. Davies J, Hebert P, Hoffman C. Royal College of Physicians and Surgeons. 2003. Canadian patient safety dictionary URL: http://www.royalcollege.ca/portal/page/portal/rc/common/documents/publications/patient_safety_dictionary_e.pdf [accessed 2014-06-03] [WebCite Cache ID 6Q5CZAmks]
4. Kohn LT, Corrigan J, Donaldson MS. To err is human: Building a safer health system. Washington, D.C: National Academy Press; 2000.
5. Baker R, Norton PG, Flintoft V, Blais R, Brown A, Cox J, et al. The Canadian adverse events study: The incidence of adverse events among hospital patients in Canada. CMAJ 2004 May 25;170(11):1678-1686 [FREE Full text] [Medline: 15159366]
6. Shep S, Cardiff K. Health Canada. Ottawa: Health Canada; 2005 Dec. Governance for patient safety: Lessons learned from non-health risk-critical high-reliability industries URL: http://theta.library.yorku.ca/uhtbin/cgisirsi/x/0/0/5?searchdata1=a2178209%7BCKEY%7D [accessed 2014-06-04] [WebCite Cache ID 6Q56K0Dib]
7. White J. Canadian Patient Safety Institute.: http://www.docstoc.com; 2007. Adverse event reporting and learning system: A review of the relevant literature URL: http://www.gov.nl.ca/ahe/submissions/05_CanadianPatientSafetyInstitute.pdf [accessed 2014-06-03] [WebCite Cache ID 6Q56gbILQ]
8. Williams SK, Osborn S. The development of the national reporting and learning system in England and Wales, 2001-2005. Med J Aust 2006 May 15;184(10 Suppl):S65-S68. [Medline: 16719740]
9. Vanderheyden LC, Northcott HC, Adair CE, McBrien-Morrison C, Meadows LM, Norton P, et al. Reports of preventable medical errors from the Alberta patient safety survey 2004. Healthc Q 2005;8 Spec No:107-114 [FREE Full text] [Medline: 16334082]
10. Simon A, Lee R, Cooke D, Lorenzetti D. Institute of Health Economics. Edmonton: Alberta Heritage Foundation for Medical Research; 2005. Institutional medical incident reporting systems: A review URL: http://www.ihe.ca/publications/

library/archived/institutional-medical-incident-reporting-systems-a-review/ [accessed 2014-05-28] [WebCite Cache ID 6PutZBJdh]
11. Eastern Health. Eastern Health Authority. St. John's: Author; 2008. Eastern health regional authority annual report: 2007-08 URL: http://www.eha.sa.gov.au/uploadFiles/listDocuments/annual_report_2008.pdf [accessed 2014-05-28] [WebCite Cache ID 6PutqQoT8]
12. Government of Newfoundland and Labrador. Newfoundland Labrador Canada. St. John's: Author; 2008. Provincial task force on adverse health events URL: http://www.gov.nl.ca/ahe/ [accessed 2014-05-29] [WebCite Cache ID 6PwQ9BID4]
13. Harris AD, McGregor JC, Perencevich EN, Furuno JP, Zhu J, Peterson DE, et al. The use and interpretation of quasi-experimental studies in medical informatics. J Am Med Inform Assoc 2006;13(1):16-23 [FREE Full text] [doi: 10.1197/jamia.M1749] [Medline: 16221933]
14. Neville D, Gates K, MacDonald D, Barron M, Tucker S, Cotton S, et al. http://www.nlchi.nl.ca. St. John's: Newfoundland and Labrador Centre for Health Information; 2004. Towards an evaluation framework for electronic health records initiatives: A proposal for an evaluation framework URL: http://www.nlchi.nl.ca/images/PDFs/final_framework.pdf [accessed 2014-05-29] [WebCite Cache ID 6PwRBozOF]
15. Delone W, McLean E. 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 2014-05-29] [WebCite Cache ID 6PwRQS1DM]
16. Lau F, Hagens S, Muttitt S. A proposed benefits evaluation framework for health information systems in Canada. Healthc Q 2007;10(1):112-116, 118. [Medline: 17326376]
17. BC Patient Safety & Learning System. Http://www.bcpsls.ca. Vancouver: British Columbia Patient Safety and Learning System; 2008. Planning the implementation of a province-wide patient safety reporting and learning system for British Columbia URL: http://www.bcpsls.ca/NR/rdonlyres/AFDEBFB3-D2BC-42BB-83D2-AD1BEC772102/50015/PSLSEvaluationReport_FINAL_Jan2508_website1.pdf [accessed 2014-05-29] [WebCite Cache ID 6Q58KFcPP]
18. Ginsburg L, Tregunno D, Fleming M, Flemons W, Gilin D, Norton P. Toronto. Perception of patient safety culture in six healthcare organizations - final report prepared from a grant from the Canadian patient safety institute URL: http://www.yorku.ca/patientsafety/psculture/reports_docs/PSC_2007_MainReport.pdf [accessed 2014-05-30] [WebCite Cache ID 6PxqNVJ8X]
19. Accreditation Canada. http://www.accreditation.ca. Ottawa: Accreditation Canada; 2012. Patient safety strategy URL: http://www.accreditation.ca/search/site/Patient%20Safety%20Strategy [accessed 2014-05-29] [WebCite Cache ID 6PwSXngtc]
20. Braithwaite J, Westbrook M, Travaglia J. Attitudes toward the large-scale implementation of an incident reporting system. Int J Qual Health Care 2008 Jun;20(3):184-191 [FREE Full text] [doi: 10.1093/intqhc/mzn004] [Medline: 18337286]
21. Braithwaite J, Westbrook MT, Travaglia JF, Hughes C. Cultural and associated enablers of, and barriers to, adverse incident reporting. Qual Saf Health Care 2010 Jun;19(3):229-233. [doi: 10.1136/qshc.2008.030213] [Medline: 20534716]
22. Cochrane D, Taylor A, Miller G, Hait V, Matsui I, Bharadwaj M, et al. Establishing a provincial patient safety and learning system: Pilot project results and lessons learned. Healthc Q 2009;12 Spec No Patient:147-153 [FREE Full text] [Medline: 19667793]
23. Frankel A, Gandhi TK, Bates D. Improving patient safety across a large integrated health care delivery system. Int J Qual Health Care 2003 Dec;15 Suppl 1:i31-i40 [FREE Full text] [Medline: 14660521]
24. Mekhjian HS, Bentley TD, Ahmad A, Marsh G. Development of a Web-based event reporting system in an academic environment. J Am Med Inform Assoc 2004;11(1):11-18 [FREE Full text] [doi: 10.1197/jamia.M1349] [Medline: 14527972]
25. Keistinen T, Kinnunen M. Increased patient safety with an Internet-based reporting system. World Hosp Health Serv 2008;44(2):37-39. [Medline: 18795505]
26. Levtzion-Korach O, Alcalai H, Orav E, Graydon-Baker E, Keohane C, Bates D, et al. Evaluation of the contributions of an electronic web-based reporting system: Enabling action. J Patient Saf 2009 Mar;5(1):9-15. [doi: 10.1097/PTS.0b013e318198dc8d] [Medline: 19920433]
27. Tepfers A, Louie H, Drouillard M. Developing an electronic incident report: Experiences of a multi-site teaching hospital. Healthc Q 2007;10(2):117-122. [Medline: 17491576]
28. Tuttle D, Holloway R, Baird T, Sheehan B, Skelton WK. Electronic reporting to improve patient safety. Qual Saf Health Care 2004 Aug;13(4):281-286 [FREE Full text] [doi: 10.1136/qhc.13.4.281] [Medline: 15289631]
29. Walsh K, Antony J. Improving patient safety and quality: What are the challenges and gaps in introducing an integrated electronic adverse incident and recording system within health care industry? Int J Health Care Qual Assur 2007;20(2-3):107-115. [Medline: 17585610]
30. Clarke JR, Lerner JC, Marella W. The role for leaders of health care organizations in patient safety. Am J Med Qual 2007;22(5):311-318. [doi: 10.1177/1062860607304743] [Medline: 17804390]
31. Sari A, Sheldon T, Cracknell A, Turnbull A. Sensitivity of routine system for reporting patient safety incidents in an NHS hospital: Retrospective patient case note review. BMJ 2007 Jan 13;334(7584):79 [FREE Full text] [doi: 10.1136/bmj.39031.507153.AE] [Medline: 17175566]
32. Kuo YH, Lee TT, Mills ME, Lin KC. The evaluation of a web-based incident reporting system. Comput Inform Nurs 2012 Jul;30(7):386-394. [doi: 10.1097/NXN.0b013e31825106ea] [Medline: 22433431]

33. Boxwala A, Dierks M, Keenan M, Jackson S, Hanscom R, Bates D, et al. Organization and representation of patient safety data: Current status and issues around generalizability and scalability. J Am Med Inform Assoc 2004;11(6):468-478 [FREE Full text] [doi: 10.1197/jamia.M1317] [Medline: 15298992]
34. Zboril-Benson LR, Magee B. How quality improvement projects influence organizational culture. Healthc Q 2005;8 Spec No:26-31 [FREE Full text] [Medline: 16334068]
35. Hirose M, Regenbogen SE, Lipsitz S, Imanaka Y, Ishizaki T, Sekimoto M, et al. Lag time in an incident reporting system at a university hospital in Japan. Qual Saf Health Care 2007 Apr;16(2):101-104 [FREE Full text] [doi: 10.1136/qshc.2006.019851] [Medline: 17403754]
36. Blais R, Bruno D, Bartlett G, Tamblyn R. Journal of Patient Safety. 2008. Can we use incident reports to detect hospital adverse events? URL: http://journals.lww.com/journalpatientsafety/Abstract/2008/03000/Can_We_Use_Incident_Reports_to_Detect_Hospital.3.aspx [accessed 2014-06-03] [WebCite Cache ID 6Q59mkoT3]
37. Milch CE, Salem DN, Pauker SG, Lundquist TG, Kumar S, Chen J. Voluntary electronic reporting of medical errors and adverse events. An analysis of 92,547 reports from 26 acute care hospitals. J Gen Intern Med 2006 Feb;21(2):165-170 [FREE Full text] [doi: 10.1111/j.1525-1497.2006.00322.x] [Medline: 16390502]

Abbreviations

BC: British Columbia
IT: information technology
QRM: quality and risk management
RN: registered nurse

Edited by G Eysenbach; submitted 11.02.14; peer-reviewed by F Ehrler; comments to author 07.04.14; revised version received 20.04.14; accepted 15.05.14; published 11.06.14.

Please cite as:
Elliott P, Martin D, Neville D
Electronic Clinical Safety Reporting System: A Benefits Evaluation
JMIR Med Inform 2014;2(1):e12
URL: http://medinform.jmir.org/2014/1/e12/
doi:10.2196/medinform.3316
PMID:25600569

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org