JMIR Medical Informatics

Clinical informatics, decision support for health professionals, electronic health records, and eHealth infrastructures
Volume 4 (2016), Issue 4 ISSN: 2291-9694

Contents

Original Papers

XSL•FO
RenderX

Viewpoint

Reviews

XSL•FO
RenderX

Original Paper

Weighting Primary Care Patient Panel Size: A Novel Electronic Health Record-Derived Measure Using Machine Learning

Alvin Rajkomar[1*], MD; Joanne Wing Lan Yim[2*], PhD; Kevin Grumbach[3,4], MD; Ami Parekh[1,3], MD, JD

[1]Department of Medicine, University of California, San Francisco, San Francisco, CA, United States

[2]Clinical Systems, University of California, San Francisco, San Francisco, CA, United States

[3]Office of Population Health and Accountable Care, UCSF Health, University of California, San Francisco, San Francisco, CA, United States

[4]Department of Family and Community Medicine, University of California, San Francisco, San Francisco, CA, United States

[*]these authors contributed equally

Corresponding Author:
Alvin Rajkomar, MD
Department of Medicine
University of California, San Francisco
533 Parnassus Ave
San Francisco, CA, 94143
United States
Phone: 1 415 476 5924
Fax: 1 415 514 2094
Email: alvin.rajkomar@ucsf.edu

Abstract

Background: Characterizing patient complexity using granular electronic health record (EHR) data regularly available to health systems is necessary to optimize primary care processes at scale.

Objective: To characterize the utilization patterns of primary care patients and create weighted panel sizes for providers based on work required to care for patients with different patterns.

Methods: We used EHR data over a 2-year period from patients empaneled to primary care clinicians in a single academic health system, including their in-person encounter history and virtual encounters such as telephonic visits, electronic messaging, and care coordination with specialists. Using a combination of decision rules and k-means clustering, we identified clusters of patients with similar health care system activity. Phenotypes with basic demographic information were used to predict future health care utilization using log-linear models. Phenotypes were also used to calculate weighted panel sizes.

Results: We identified 7 primary care utilization phenotypes, which were characterized by various combinations of primary care and specialty usage and were deemed clinically distinct by primary care physicians. These phenotypes, combined with age-sex and primary payer variables, predicted future primary care utilization with R^2 of .394 and were used to create weighted panel sizes.

Conclusions: Individual patients' health care utilization may be useful for classifying patients by primary care work effort and for predicting future primary care usage.

(JMIR Med Inform 2016;4(4):e29) doi:10.2196/medinform.6530

KEYWORDS

primary health care; risk adjustment; patient acceptance of health care; ambulatory care; health care economics and organizations; medical informatics; machine learning

Introduction

In the face of increasing demand for primary care services [1] and concerns of a primary care physician (PCP) shortage [2], health systems need methods to effectively match primary care workload and capacity [3]. Empanelment, assigning each patient to a primary care physician (PCP) or team, is an essential building block for high-performing primary care [4,5].

Health systems moving toward empaneled models of care must account for the truism that no two patients are the same; different patients require substantially different amounts of primary care

XSL•FO
RenderX

work effort to address their health care needs [3]. Methods are needed to acknowledge and predict how much primary care work effort a patient needs in order to adjust panel sizes to account for differences in patient mix across individual PCPs and practices to better match capacity with demand. The methods could also be used to adjust panel-based payment to pay a higher capitated rate for patients requiring more primary care work effort.

Traditional methods to adjust panel size using basic patient demographic data such as age and sex have limited predictive power [6]. These approaches have been augmented by other approaches that are limited by requiring multiple data sources (eg, pharmacy data and insurance claims), poor utility in predicting primary care work effort, their proprietary natures, and lack of validation in the literature [7-11].

The lack of a validated predictive model and the desire of our academic health system to use case-mix–adjusted primary care physician (PCP) panel sizes in our own operations motivated us to use machine learning methodologies on regularly collected electronic health record (EHR) data to create a novel method to adjust panel sizes. Given the variety of diagnoses possible in a population and the spectrum of care complexity for different patients with the same diagnoses, the phenotypes in our model are based on objectively measured interactions with the health system rather than on disease-based codes entered by clinicians in the EHR. In this paper, we describe our method of using patient utilization phenotypes to better characterize primary care work effort to develop a novel methodology for weighting primary care panel size.

Methods

Overview of Study Design

Our overall study design consisted of 3 major steps.

Define utilization phenotype clusters: We used a training set sample of patients with year 1 EHR data on health system encounters to cluster patients into distinct utilization phenotypes, using k-means clustering methods.

Validate utilization phenotype clusters: We determined among patients in a separate test set if models using utilization phenotype clusters were better at predicting year 2 primary care visits than models using simpler, raw counts of year 1 encounters, using log-linear regression models.

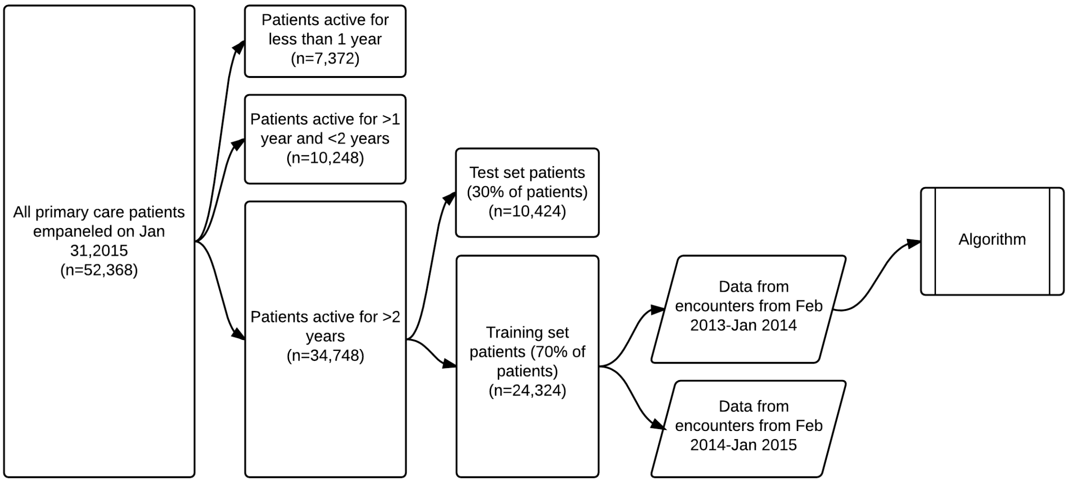
Determine weights for each phenotype cluster for computing weighted panel sizes: We consolidated utilization phenotype clusters into a smaller number of final primary care work clusters, weighted each final cluster based on median number of concurrent year primary care visits among patients in each cluster, and applied these weights to the entire sample of empaneled patients.

Sample and Data Sources

We used the EHR system (Epic, Madison, WI, USA) to collect data on all patients older than 18 years empaneled as of January 31, 2015, to a primary care clinician in practices operated by the University of California, San Francisco health system (UCSF Health). Empanelment at UCSF Health is defined as having an identified UCSF Health primary care clinician listed in the EHR primary care provider field and at least 1 visit in the prior 3 years to any clinician at the primary care practice; 52,368 adult patients were empaneled at primary care practices in January 2015.

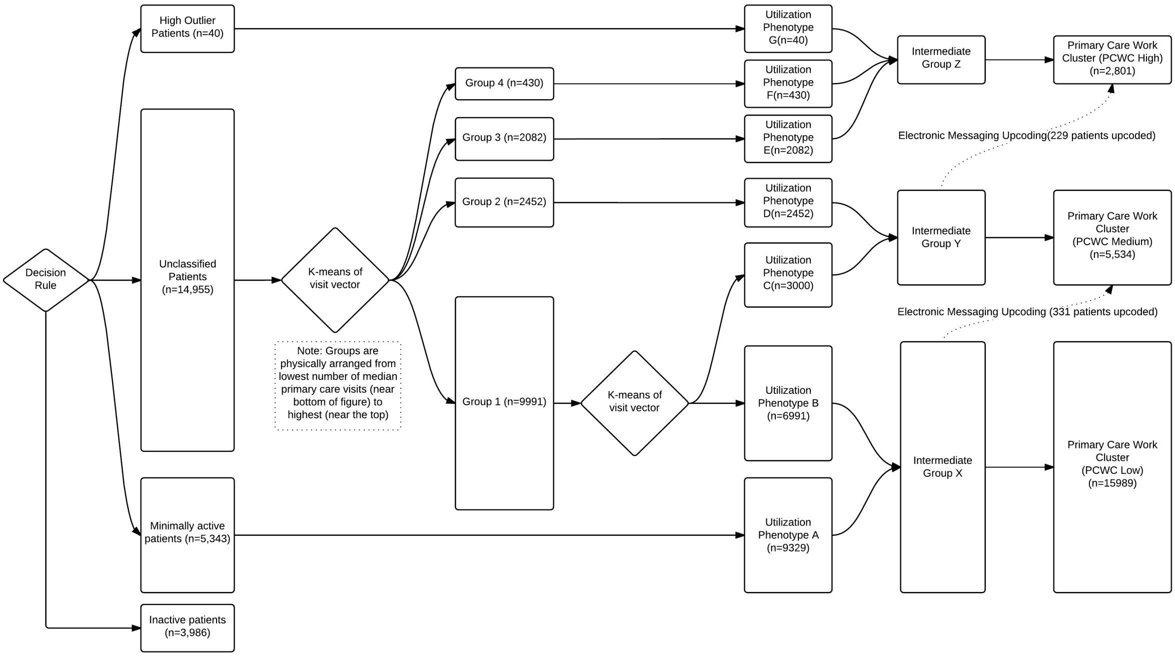
For model development, we included only the subset of 34,748 patients who had at least 1 encounter (including office visit, telephone, electronic messaging, or medication refill) occurring on or before February 1, 2013, to ensure that patients in the study would have at least 12 months of eligibility for data analysis for deriving the predictive model (February 1, 2013 to January 31, 2014) and then a subsequent 12 months of data for using the model to predict utilization (February 1, 2014 to January 31, 2015). The model was developed on a training set of a random sample of 70.00% (24,324/34,748) of these patients, and the remaining 30.00% (10,424/34,748) were left as a test set (Figures 1 and 2).

Figure 1. Flowchart of data from the electronic health record to the algorithm. PCWC: Primary care work cluster.


All primary care patients empaneled on Jan 31,2015 (n=52,368)
Patients active for less than 1 year (n=7,372)
Patients active for >1 year and <2 years (n=10,248)
Patients active for >2 years (n=34,748)
Test set patients (30% of patients) (n=10,424)
Training set patients (70% of patients) (n=24,324)
Data from encounters from Feb 2013-Jan 2014
Data from encounters from Feb 2014-Jan 2015
Algorithm

XSL•FO
RenderX

Figure 2. Flowchart of decision rules and clustering algorithms that demonstrate how patients were classified into different utilization phenotypes and primary work group clusters.


Decision Rule
High Outlier Patients (n=40)
Unclassified Patients (n=14,955)
Minimally active patients (n=5,343)
Inactive patients (n=3,986)
K-means of visit vector
Note: Groups are physically arranged from lowest number of median primary care visits (near bottom of figure) to highest (near the top)
Group 4 (n=430)
Group 3 (n=2082)
Group 2 (n=2452)
Group 1 (n=9991)
K-means of visit vector
Utilization Phenotype G(n=40)
Utilization Phenotype F(n=430)
Utilization Phenotype E(n=2082)
Utilization Phenotype D(n=2452)
Utilization Phenotype C(n=3000)
Utilization Phenotype B (n=6991)
Utilization Phenotype A (n=9329)
Intermediate Group Z
Intermediate Group Y
Intermediate Group X
Primary Care Work Cluster (PCWC High) (n=2,801)
Primary Care Work Cluster (PCWC Medium) (n=5,534)
Primary Care Work Cluster (PCWC Low) (n=15989)
Electronic Messaging Upcoding(229 patients upcoded)
Electronic Messaging Upcoding (331 patients upcoded)

Variables Used

Variables Included in Weighting Algorithm

For each patient, we retrospectively collected the data for the following types of encounters at our health system between July 1, 2012 and January 31, 2015: primary care office visits billed for more than 5 minutes, missed appointments, emergency department visits, emergent hospitalizations, elective hospitalizations, infusion and transfusion center visits, medical and surgical subspecialty visits, diagnostic and interventional radiology visits, telephone encounters with any member of their assigned primary care team, urgent care visits, and electronic messages with their primary care team through the EHR secure messaging system. In addition, we collected demographic data including age, sex, race-ethnicity, primary payer, primary care clinic location, and primary care clinician. We also included every medication documented in the EHR medication list, including start and stop dates.

For each patient in the training set, we created a visit vector that represented the various encounters across the health system. Each component of the vector was created by summing the total number of visits within a respective encounter type that occurred from February 1, 2013 to January 31, 2014. The encounter types included "effective number" of primary care visits (an adjusted visit count incorporating medication counts, as defined below), telephone encounters with the primary care office, missed appointments to the primary care office, urgent care visits, emergency room visits, emergent hospitalizations, routine hospitalizations, medical and surgical specialty visits, infusion center visits, transfusion center visits, diagnostic and interventional radiology visits, and electronic messaging. Other than primary care and specialty visits, each encounter was an equal contributor to its respective category.

We created an effective number of primary care visits to account for additional time required for medication reconciliation and complexity of PCP visits for patients with multiple medications. For each visit, we calculated the number of active medications. If there were 5 or fewer active medications at that particular primary care visit, then that visit was assigned a weight of 1. If there were 6 to 10 medications, then the visit was assigned a weight of 1.5. If the medication count was greater than 10, then the visit was assigned a weight of 1.75. The "effective" primary care office visit count was the sum of these weighted visits.

Some specialists care for diseases that require frequent visits, such as weekly dermatologic treatments, and other specialists may often monitor diseases that require only yearly follow-ups. Because of the high standard deviation of visit counts per year for different specialties, we capped the total number of visits counted for each specialty. The cap was set for each specialty separately at 2 standard deviations above the mean number of visits per year among all patients seen by that specialty. For example, if a patient had 20 dermatology visits and 2 cardiology visits, the total number of specialty visits we counted was 17.8 because the cap for dermatology visits was 15.8 and for cardiology visits 6.7.

Additional Variables Included in Algorithm Validation

For validation of the algorithm, we also included age, sex, and primary payer. Patients were split by age and sex into 12 categories, using the age groups 18-34, 35-49, 50-64, 65-69, 70-84, and 85-115 years. The 3 patients with missing sex were categorized as female in order to keep the patients in the analytic

XSL•FO
RenderX

sample. The primary payers were characterized as commercial, Medicare, Medicaid, or other.

Primary Care Focus Group for Expert Consensus

As we refined the algorithm, we asked a focus group of practicing PCPs to qualitatively evaluate whether the clusters our methodology identified aligned with their perception of the level of work needed for their patients. The group included 15 family physicians and general internists.

Algorithm

The algorithm was developed using only the patients in the training set. We used a decision rule for initial classification of patients (Figure 3). Patients with greater than 6 standard deviations above the mean number of annual primary care visits were classified as “high outliers.” Patients who met all the following criteria were classified as “minimally active”: ≤1 primary care visit per year, 0 emergency department visits, 0 hospitalizations, ≤4 specialty visits per year, ≤2 telephone encounters per year, and ≤6 electronic messages to the patient per year. However, if patients had zero visits across all these categories (excluding missed appointments), then they were classified as “inactive patients.” Only patients not meeting the criteria for “high outliers,” “minimally active,” and “inactive patients” entered the next stage of the algorithm. In the algorithm, these patients were divided into 4 groups by k-means clustering on the encounter vectors. At this point, all variables were of the same unit of analysis (eg, number of visits per year), which made the clusters easier to interpret. The selective truncation of some of the visit types as described in the Variables Included in Weighting Algorithms section was utilized in place of blindly normalizing by mean and standard deviation. All encounter categories except for electronic messaging were used in this step. The k-means clustering was performed using the Hartigan-Wong algorithm. We used 4 centers with 5 random initiations and up to a maximum of 10 iterations to find stable cluster definitions. We chose to use 4 clusters by examining the change in reduction of the within-group sum of squares (Multimedia Appendix 1) and by verifying with clinicians that their own patients assigned to the clusters were meaningfully distributed (see below).

The clusters were then ranked by the median annual number of raw PCP visits (ie, visit counts that were not weighted for number of medications). Our primary care physician (PCP) focus group decided that the cluster with the fewest visits contained 2 heterogeneous groups after examining the assignments of their own patients. Therefore, that cluster was further divided in 2 by k-means clustering, which aside from the number of clusters used the same algorithm and settings as the previous clustering (Figure 3).

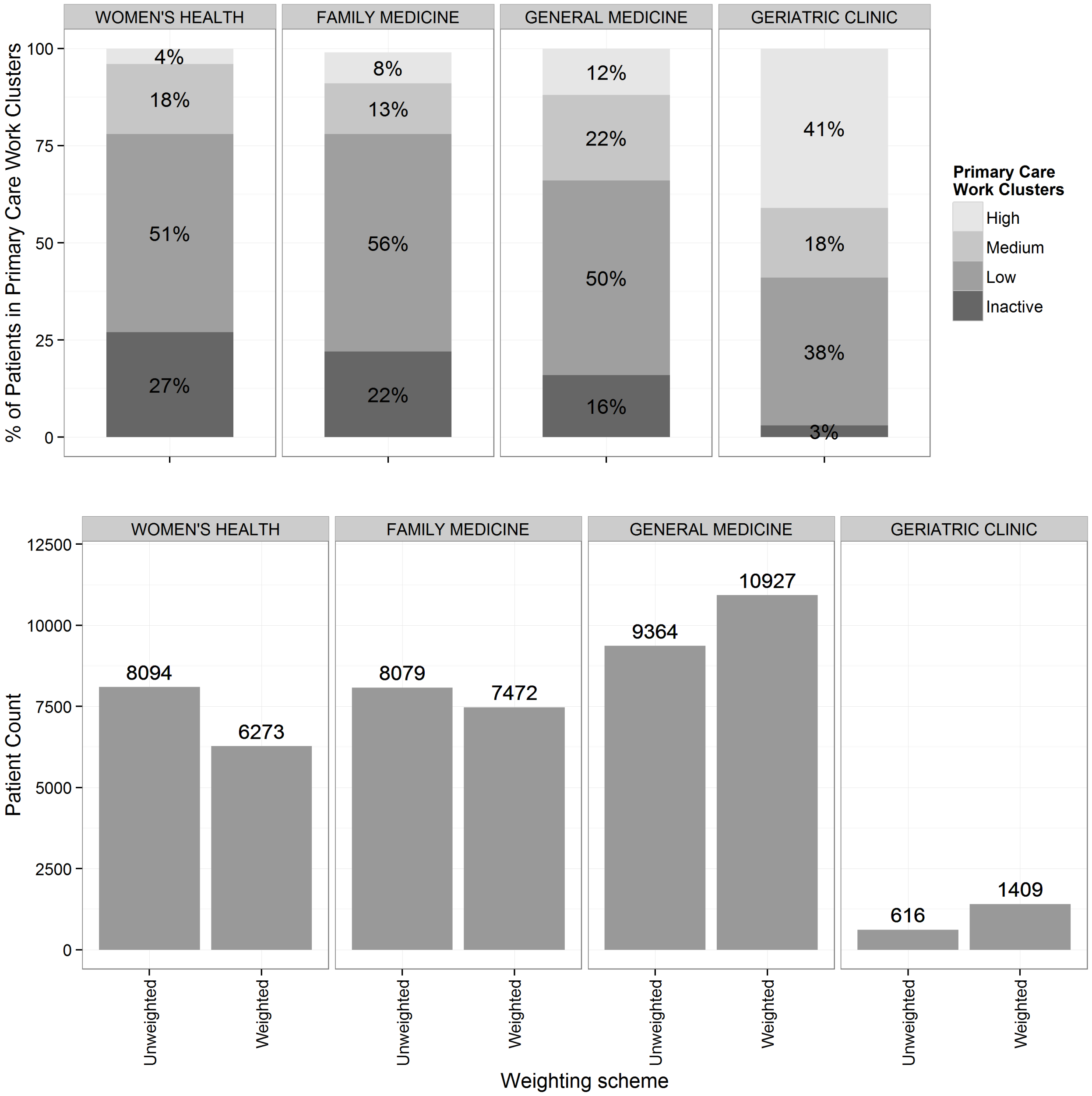
Excluding the inactive patients, there were 7 resulting groups: 2 from the initial decision rules, 3 from the initial cluster assignment, and 2 from the second round. These 7 cluster groups represented different patterns of health care utilization across the health system—health care utilization phenotypes—which we labeled A through G (Figure 3).

The focus group of PCPs agreed that the groupings represented distinct primary care phenotypes but believed that some of the phenotypes required a similar amount of primary care work effort. Therefore, we collapsed the 7 phenotypes into 3 categories—intermediate groups X, Y, and Z (Figure 3), ranked by the median number of primary care visits per year among patients in the group.

A final decision rule was applied to account for patients’ use of secure electronic messaging with their providers. Patients who sent more than 1.5 standard deviations of electronic messages relative to the mean of all patients in the originally assigned category or who were sent more than 24 messages by their primary care clinician were moved to the next higher cluster. The final clusters were labeled high, medium, and low to represent the relative amount of primary work effort for patients in that cluster, with a fourth cluster being the inactive patients. Patients initially classified as “minimally active” were added to the “low” group. We refer to these as primary care work clusters.

XSL•FO
RenderX

Figure 3. The fractions of all patients assigned to the primary care work clusters in 4 selected clinics and their unweighted and weighted panel sizes. The distribution of patients across clusters was unique to each clinic, and because each cluster is weighted differently, the difference between weighted and unweighted panel sizes differed for each clinic as well. The geriatric clinic, which has 41% of its population assigned to the high work cluster, had a weighted panel size that was more than twice the unweighted size.



Validation

The utilization phenotypes were designed to cluster patients based on utilization patterns in a nonhypothesis-driven way. To demonstrate that the clusters had predictive power, we sought to validate them as part of a risk adjustment model predicting subsequent primary care service utilization. We created a series of generalized linear models to predict the total number of primary care encounters (PCP visits and telephone encounters) for each patient in the second year (February 1, 2014 to January 31, 2015). The models were developed using the same patients in the training set sample. As predictors, we used age-sex categories, payer type, and one of two variables measuring utilization patterns during the first year (February 1, 2013 to January 31, 2014): the 7 primary care utilization phenotypes (which we have described above) or a simpler measure of the raw counts of all types of encounters (which we refer to as the “naïve phenotype”). We used the 7 utilization phenotypes rather than the 3 work clusters, which are derived from the phenotypes, because the phenotypes were felt to encode meaningful clinical distinctions by the primary care focus group. The naïve phenotype was created by summing the total number of all in-person encounters (primary care visits, emergency department visits, hospitalizations, infusion and transfusion visits, urgent care visits, specialty visits, and outpatient procedures for cardiology, radiology, pulmonology, and neurology). These sums were rank ordered and divided into 7 percentiles so as to have the same number of categories as the primary care work clusters.

XSL•FO
RenderX

We then applied the coefficients derived from the training set to predict the log number of primary care visits in the second year for the test set of the sample, which was not used to generate the model. We report the adjusted *R*-squared and the Akaike information criterion (AIC). The AIC is a goodness-of-fit value that balances model bias versus variability, ranges from 0 to infinity, and penalizes models with more variables.

We repeated the analysis with the outcome of the number of primary care visits only (not including telephone encounters). We also repeated the analysis modeling the raw rather than log number of visits per year with a Poisson and zero-inflated Poisson distribution with a canonical log link.

Weighted Panel Size

The work clusters were used to calculate weighted panel sizes as of February 1, 2015, using all 52,368 adult patients empaneled in primary care. We assigned patients to 4 primary care work clusters using the algorithm defined by the training set, as described above, based on EHR data on activity at our health system between February 1, 2014 and January 31, 2015. For patients with less than 12 months of activity, we initially weighted the number of visits by the number of months the patient had an active status, but this gave patients with just a few visits with a short exposure time high counts in their visit vector (eg, 2 visits in 3 months would be calculated to an average of 8 visits per year). Instead, for those patients we assumed their visits were over 12 months.

Once patients were assigned to a primary care work cluster, we needed to assign weights to each of the 4 final clusters (high, medium, low, and inactive). In consultation with our focus group of clinicians, we decided to base the weights on the number of effective primary care visits among patients in each of the clusters between February 1, 2014 and January 31, 2015. The relative weights of the “medium” and “high” clusters were defined by dividing the median number of effective primary care visits among patients in each of these clusters by the median number of effective primary care visits in the “low” cluster. Because patients in the “inactive” cluster had no activity in the preceding 12 months but were still empaneled in primary care and might be expected to have some future activity, we assigned patients in the inactive cluster a weight of 0.05.

Finally, to make the total number of weighted patients equal the total number of raw, unweighted patients empaneled in primary care (ie, 52,368), we used an additional scaling factor, *w*, to impose this restriction. (Figure 4)

The cluster weights for the low, medium, and high clusters were then defined to be *w* multiplied by the median number of PCP visits of the respective cluster divided by the median number of patients in the low cluster (Figure 4). To calculate an effective panel size for a clinic or primary care provider, each patient in the panel was classified to a primary care work cluster. The number of patients in each cluster was multiplied by the respective weight, and the sum over all clusters defined the weighted panel size.

To demonstrate how panel sizes for PCPs changed from the raw panel size to the weighted panel size, we calculated the average change in panel size. In this analysis, we only included PCPs who had an unweighted panel size of greater than 150 active patients.

All analyses were performed using R version 3.1.2 (R Foundation for Statistical Computing). The k-means algorithm was from the standard “stats” package (version 3.2.1). The research was approved by the Institutional Review Board at UCSF.

Figure 4. Equations that define how scaling factor *w* was defined. We constrain the total weighted population size (the right hand side) to be equal to the total unweighted population size in (a). We solve for *w* in (b) PCP: primary care physician.

(a)

$$N_{total\,population} = \sum_{i\in\{,l,m,h\}} (w \cdot X_i \cdot N_i) + 0.05 \cdot N_{inactive}$$

(b)

$$w = \frac{N_{total\,population} - 0.05 \cdot N_{inactive}}{\sum_{i\in\{l,m,h\}} X_i N_i}$$

$$N_i \overset{\text{def}}{=} number\ of\ patients\ in\ cluster\ i$$

$$P_l = median\ number\ of\ PCP\ visits\ of\ cluster\ low$$

$$X_l = 1$$

$$X_m = \frac{median\ number\ of\ PCP\ visits\ of\ cluster\ medium}{P_l}$$

$$X_h = \frac{median\ number\ of\ PCP\ visits\ of\ cluster\ high}{P_l}$$

XSL•FO
RenderX

Results

Description of the Utilization Phenotypes and Primary Care Work Clusters

Of the 52,368 adult patients empaneled on January 31, 2015, a total of 34,748 were active for more than 2 years. Those were further subdivided into training and test sets of 24,324 and 10,424 patients (Figures 1 and 2). Characteristics of the patients in the training set and their utilization are presented in Tables 1 and 2.

Of the patients in the training set, 3986 were determined to be inactive, 5343 minimally active, and 40 high-outlier patients. The remaining 14,955 patients were k-means clustered based on the visit vector into 5 utilization phenotypes. These phenotypes were combined with minimally active and high-outlier patients into 7 phenotypes, which were further merged into 3 primary care work clusters (Figure 3).

The characteristics of patients in each utilization phenotype are presented in Tables 1 and 2 (Full table is in Multimedia Appendix 2). None of the demographic variables demonstrated a monotonic increase or decrease across the phenotypes, although phenotypes E-G tended to represent older, female, patients with government health plans.

Table 1. Patient characteristics of each utilization phenotypes (inactive through group D) in the training set (N=24,324).

Characteristics	Utilization phenotype				
	Inactive	A	B	C	D
Size of group (n)	3986	5343	6991	3000	2452
Age, years, mean (SD)	41.9 (17.3)	47.7 (14.7)	53.7 (16.8)	56.6 (16.4)	59.9 (17.3)
Male, n (%)	1551 (38.9)	2057 (38.5)	2678 (38.3)	1083 (36.1)	922 (37.6)
White, n (%)	1814 (45.5)	2875 (53.8)	3293 (47.1)	1635 (54.5)	1324 (54)
Asian, n (%)	694 (17.4)	1095 (20.5)	1734 (24.8)	675 (22.5)	596 (24.3)
Black, n (%)	379 (9.5)	289 (5.4)	587 (8.4)	222 (7.4)	184 (7.5)
Commercial, n (%)	2738 (68.7)	4266 (79.9)	4348 (62.2)	1731 (57.7)	1113 (45.4)
Medicare or Medicaid, n (%)	1068 (26.8)	992 (18.6)	2545 (36.4)	1245 (41.5)	1324 (54.0)
Other payer, n (%)	180 (4.5)	85 (2)	98 (1)	24 (1)	15 (1)
Active medications at PCP[a] visit, mean (SD)	0 (0)	2.3 (2.9)	5 (3.6)	5.5 (4.3)	8.1 (6)
Primary care visits, mean (SD)	0 (0)	0.7 (0.5)	2.6 (1.3)	2.1 (1.4)	2.9 (2.3)
Weighted primary care visits, mean (SD)	0 (0)	0.7 (0.6)	3.2 (1.8)	2.8 (1.9)	4.3 (3.7)
No-show visits, mean (SD)	0.1 (0.4)	0.2 (0.7)	0.5 (1)	0.6 (1.2)	1.4 (2.1)
Urgent care visits, mean (SD)	0 (0)	0.1 (0.5)	0.2 (0.5)	0.2 (0.6)	0.2 (0.6)
Telephone encounters, mean (SD)	0 (0)	0.4 (0.7)	1.7 (1.8)	1.4 (1.6)	2.3 (2.5)
Emergency department visits, mean (SD)	0 (0)	0 (0)	0.2 (0.5)	0.2 (0.5)	0.3 (0.7)
Emergent hospitalizations, mean (SD)	0 (0)	0 (0)	0 (0.2)	0 (0.3)	0.1 (0.5)
Elective hospitalizations, mean (SD)	0 (0)	0 (0)	0 (0)	0 (0.2)	0.1 (0.3)
Specialist visits (capped), mean (SD)	0 (0)	1 (1.2)	1 (1)	5.5 (1.4)	14 (5.3)
Infusion visits, mean (SD)	0 (0)	0 (0.5)	0 (0.4)	0.1 (1)	0.7 (4.1)
Transfusion visits, mean (SD)	0 (0)	0 (0.8)	0 (0.2)	0.1 (0.8)	0.4 (2.6)
Radiology or procedure visits, mean (SD)	0 (0)	0.4 (0.8)	0.6 (1)	1.2 (1.4)	2.2 (2.6)
Secure electronic messages to patient, mean (SD)	0 (0)	0.7 (1.4)	2.3 (4.3)	4 (6.4)	6.8 (11.1)
Secure electronic messages from patient, mean (SD)	0 (0)	0.9 (1.8)	2.8 (5.3)	5 (8.2)	8.9 (15)

[a] PCP: primary care physician.

XSL•FO
RenderX

Table 2. Patient characteristics of each utilization phenotype (group E to G) in the training set (N=24,324). The total column includes data from phenotypes in Table 1.

Characteristics	Utilization phenotype			
	E	F	G	Total sample
Size of group (n)	2082	430	40	24,324
Age, years, mean (SD)	65.1 (16.8)	67.4 (16.3)	60.5 (14.4)	52.7 (17.9)
Male, n (%)	716 (34.4)	158 (36.7)	8 (20)	9170 (37.7)
White, n (%)	799 (38.4)	191 (44.4)	14 (35)	11,943 (49.1)
Asian, n (%)	525 (25.2)	76 (18)	4 (10)	5400 (22.2)
Black, n (%)	385 (18.5)	102 (23.7)	17 (43)	2165 (8.9)
Commercial, n (%)	431 (20.7)	26 (6)	3 (8)	14,665 (60.3)
Medicare or Medicaid, n (%)	1628 (78.2)	402 (93.5)	37 (93)	9219 (38.0)
Other payer, n (%)	23 (1)	2 (1)	N/A[a]	440 (1.8)
Active medications at PCP[b] visit, mean (SD)	11 (5)	15.7 (6.1)	16.2 (9.3)	4.7 (5.2)
Primary care visits, mean (SD)	7 (2.8)	11.5 (4.5)	33.2 (10)	2.3 (3)
Weighted primary care visits, mean (SD)	10.7 (4.4)	19.1 (7.8)	53.1 (15.6)	3.2 (4.7)
No-show visits, mean (SD)	1.8 (2.4)	4.3 (4.9)	6.2 (5.3)	0.7 (1.6)
Urgent care visits, mean (SD)	0.2 (0.7)	0.5 (1.2)	1.2 (2.4)	0.1 (0.5)
Telephone encounters, mean (SD)	5.9 (3.8)	19.4 (10.3)	18.5 (22.5)	1.9 (3.8)
Emergency department visits, mean (SD)	0.5 (1)	1.6 (2.9)	1.8 (2.1)	0.2 (0.7)
Emergent hospitalizations, mean (SD)	0.2 (0.5)	0.9 (1.7)	0.9 (1.5)	0.1 (0.4)
Elective hospitalizations, mean (SD)	0 (0.2)	0.1 (0.4)	0 (0.2)	0 (0.1)
Specialist visits (capped), mean (SD)	4.4 (3.3)	11.5 (8.2)	7.6 (9.2)	3.2 (4.9)
Infusion visits, mean (SD)	0.1 (2.2)	0.1 (1.1)	0 (0.2)	0.1 (1.5)
Transfusion visits, mean (SD)	0 (0.6)	0.5 (3.3)	0.2 (1.3)	0.1 (1.1)
Radiology or procedure visits, mean (SD)	1.5 (1.7)	2.7 (3)	2.5 (2.9)	0.8 (1.5)
Secure electronic messages to patient, mean (SD)	3.4 (7.8)	5.6 (14.6)	5 (14.5)	2.4 (6.1)
Secure electronic messages from patient, mean (SD)	4.4 (10.4)	8.1 (22.3)	8.6 (25.5)	3.1 (8.1)

[a] N/A: not applicable.

[b] PCP: primary care physician.

Patients with utilization phenotype A saw their primary care physician (PCP) less than once a year and tended not to have much health care exposure across the health system. Patients with phenotypes B, C, and D had a mean of 2 or more visits a year with their PCPs, although those with C had more than 5 times the average number of specialty visits compared with those with B, and D had 14 times more. Phenotypes E and F saw their PCPs on average more than 7 times a year, with phenotype F also having more than double the number of specialty visits compared with E. Phenotype G was predefined as the “high-outlier” group and saw their primary care doctor on average more than 30 times a year.

The characteristics of patients in each of the final 4 primary care work clusters, which were created by combining utilization phenotypes, are provided in Multimedia Appendix 3. The medium and high clusters tended to be older and female and to have Medicare or Medicaid. There was a monotonic increase of nearly every component of the encounter vector except for specialist visits and infusion center visits.

Validation of the Prediction of Primary Care Office and Telephone Visits

The results of the linear models to predict log-transformed primary care office and telephone visits are presented in Table 3. A model with only age-sex and payer accounted for 20.9% of the variance of primary care office and telephone visits the next year. When we added the naïve phenotype, or groupings based on the raw number of total in-person health care encounters as described above, 34.4% of the variance was captured by the model. If the utilization phenotype was used instead of the naïve phenotype, 39.4% of the variance was modeled. This model had the lowest AIC, which indicates a better fit even accounting for additional variables in the model. The results were similar with generalized linear models of a

XSL•FO
RenderX

zero-inflated Poisson regression predicting unlogged counts, which are not shown.

We also report the results of a similar model predicting just the office-based primary care visits (Multimedia Appendix 4) where the age-sex group, payer, and utilization phenotype demonstrated the best fit of the data with 34.4% of the variance of log of visit number captured.

Weighted Panel Sizes

Using the entire sample, we calculated the weights for the different primary care work clusters using concurrent year primary care visits to determine weights, as described above. The inactive, low, medium, and high clusters had weights of 0.050, 0.659, 1.319, and 4.396, respectively.

Whereas the unweighted sizes of the inactive and low clusters were 11,830 and 26,091, the weighted sizes of these populations decreased to 591 and 17,205, respectively. Conversely, the weighted sizes of the medium and high clusters increased from 9404 to 12,402 and from 5043 to 22,169, respectively (Multimedia Appendix 5). By definition, the total unweighted population size was equal to the weighted population size.

Different clinics and PCPs had different proportions of patients in high, medium, and low clusters. Illustrative results for 4 primary care clinics caring for adults are displayed in Figure 3. Patients in the high and medium clusters combined constituted slightly more than 20% of adult patients at the Women's Health Primary Care and Family Medicine Clinics, compared with 34% of patients in the General Medicine Clinic and 59% of patients in the Geriatric Clinic. Correspondingly, weighted adult panel sizes were smaller than unweighted raw panel sizes in Women's Health Primary Care and Family Medicine Clinics (decreasing in size from 8094 to 6273 and 8079 to 7472, respectively), whereas the weighted panel size was somewhat greater than unweighted at the General Medicine Clinic (9364 unweighted and 10927 weighted) and more than twice as large at the Geriatric Clinic (616 unweighted and 1409 weighted).

The relative change in panel size for each individual primary care physician (PCP) is displayed in Figure 5. The relative change in panel size between weighted and unweighted ranged from a relative decrease of 50% to a relative increase of 150%. Two physicians, who care for complex geriatric patients, had weighted panel sizes that were more than double their raw panel sizes. A total of 52% of physicians had a relative change in panel size of 20% or less. Using individual physicians as the unit of analysis, the mean weighted panel size of panel sizes greater than 150 was 12.8% greater than the mean unweighted panel size. The mean change including all panel sizes is 0. Panel sizes that were less than 150 were usually due to physicians working fewer sessions per week (and therefore caring for fewer patients).

Table 3. Log-linear model using demographic variables and baseline utilization phenotype to predict subsequent year primary care telephone encounters and office visits among patients in the test set.

Model predictors	Adjusted R^2	AIC[a]
Age-sex[b]	.166	60,780
Payer[c]	.128	61,495
Naïve phenotypes (NP)[d]	.259	57,724
Primary care cluster utilization phenotype (UP)[e]	.330	55,088
Age-sex and payer	.209	59,450
Age-sex, payer, and NP	.343	54,813
Age-sex, payer, and UP	.394	52,769

[a] AIC: Akaike information criterion.

[b] Age-sex bins are categorical variables of the combination of male or female with the following age groups: 18-34, 35-49, 50-64, 65-69, 70-84, and 85-115 years.

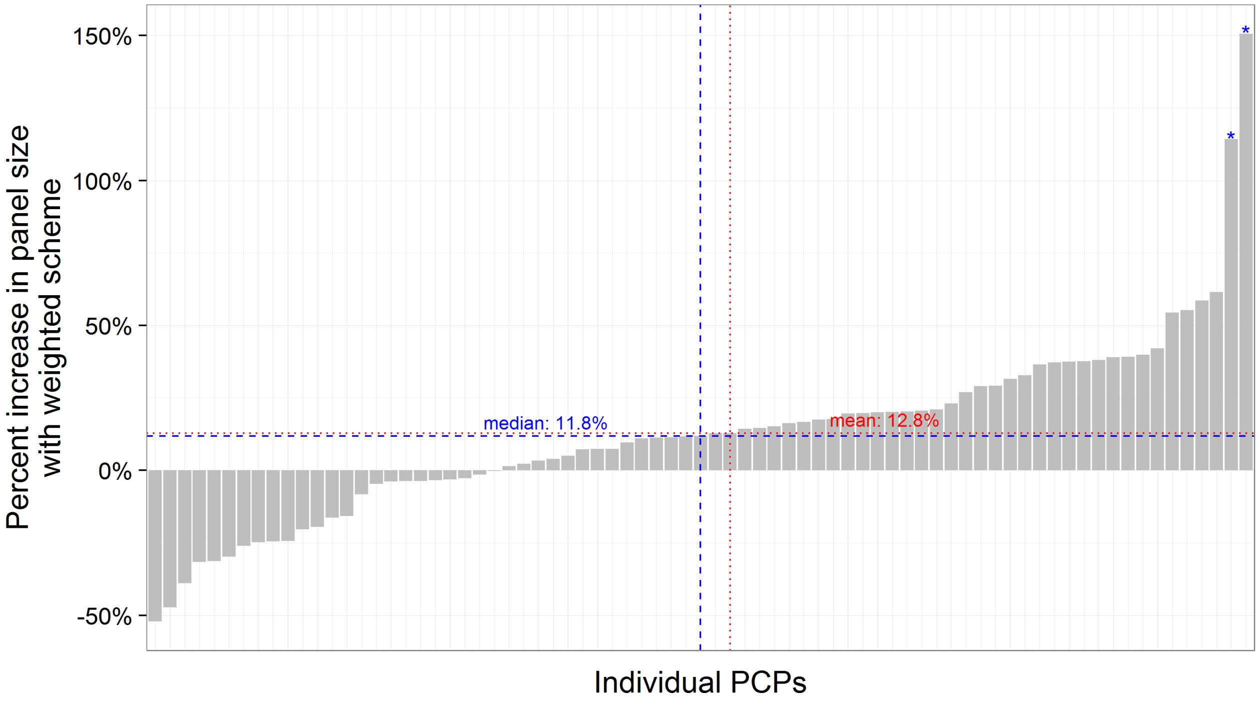
[c] Payers are defined as commercial, Medicare or Medicaid, or other.

[d] The naïve phenotype is a categorical variable that is obtained by summing the total number of health care encounters in the baseline year. These values were rank ordered and divided into 7 percentiles.

[e] The utilization phenotype is a categorical variable encoding 1 of the 7 phenotype clusters created by our algorithm.

XSL•FO
RenderX

Figure 5. The change of weighted panel size for various primary care providers with more than 150 patients. The panel size increases on average by 12.8%. *These 2 primary care physician (PCPs), who are geriatricians, had a panel size increase of more than 100%.



Discussion

Principal Findings

We have described a novel method of using EHR data collected as part of routine care to cluster primary care patients into groups that reflect differences in the primary care work effort required to care for diverse patients. We have demonstrated how this utilization phenotype method can be used to compute weighted panel sizes at the clinic and individual primary care physician (PCP) levels and, by inference, the relative capacity of clinics and PCPs to care for a panel of patients. The utilization phenotype method performed better than other methods in predicting primary care visits in the subsequent year and resulted in weighted panel sizes that differed from unweighted panel sizes at the clinic and individual primary care physician (PCP) levels. The weighting method had face validity when vetted among primary care clinicians caring for patients in the study sample and when comparing results for family medicine, general internal medicine, and geriatrics primary care clinics.

What are the advantages, limitations, and utility of our weighting method? One major strength is that all the data for the algorithm are routinely collected in the EHR. The method takes advantage of the "big data" opportunity afforded by EHRs to use a much richer variety and amount of data to compute weights, compared with traditional methods that primarily rely on a few data elements such as patient demographics and diagnostic codes. All the calculations are transparent (eg, one can inspect the characteristics of patients in each phenotype) and can be rerun easily. The model allows flexibility in assignment of final weights; we assigned weights based on the median number of primary care office visits in a cluster, but the weights could also be determined by consensus or expert opinion or by measurement of median primary care visits for the same clusters in a different health system. Patients are profiled on a single standard, allowing panel sizes to be compared across physicians who care for different populations, such as a geriatrician or a family physician.

We believe that our utilization phenotype approach has conceptual advantages over weighting models that rely on diagnoses coded in EHRs or insurance claims. Our approach does not assume that all patients with a similar diagnosis profile will have similar demands on a health system; instead, a patient's own activity generates a personalized profile. This allows patients with different disease states, such as interstitial lung disease, obscure gastrointestinal bleeds, or anxiety disorder, to be compared on a single, standardized scale. Reforms in diagnostic coding conventions such as the *International Classification of Diseases, Tenth Revision*, will continue to lack sufficient sensitivity in design and reliability in application to fully capture variation in disease states within diagnostic codes that are meaningful for panel weighting. Moreover, as disease becomes more active or quiescent, the dynamic changes can be reflected in the utilization phenotypes in near real-time. Patients who may have severe diseases but avoid care or use the services of other health care systems are reflected as inactive patients and not weighted highly. If they reengage in care, the new activity will then be reflected in a utilization phenotype. Patients with chronic pain and psychosocial comorbidities often require more frequent touches with the health system than their formal diagnoses would suggest. Rather than inferring primary care work demand from patients' demographics and diagnoses, our method attempts to more directly estimate work effort. We also captured measures of patient activity that may not be billed, such as secure electronic messaging, medication reconciliation, and care coordination among multiple specialists.

XSL•FO
RenderX

Limitations

There are several limitations to our algorithm. Our method accepts that observed patterns of service activity reasonably approximate patient demand for primary care work effort. The measure does not distinguish between medically necessary and unnecessary visits, telephone calls, referrals, and other services. A physician who induces inappropriate demand for services would appear in our model to have more complex patients than would a physician who avoids unnecessary services in caring for the same group of patients. However, any system that attempts to measure patient complexity can be gamed, with upcoding diagnostic assessments being a well-known liability for diagnosis-based case mix adjustment methods [12]. To intentionally increase a patient's complexity by our algorithm, a physician would have to spend more time in care activities, which carries a high opportunity cost. Health systems might consider complementing our panel weighting method with use of other methods to monitor physicians for patterns of wasteful care.

Another important limitation is that because the panel weighting is normalized within our system to make the total weighted patient count equal to the unweighted count, the method cannot be easily used to compare the relative primary care work demand of primary care patients in our system with that of patients in another system. If additional systems using the same EHR vendor begin to use this model and are willing to collaborate on the final weighting steps, cross-system comparisons may be possible. Our model also does not answer the question of what the "right" weighted panel size should be for a given health system. A final limitation is that our method does not as yet include children. We are developing a similar algorithm to apply to this population.

Conclusions

Our panel weighting model may be useful when implementing a variety of health system policies related to primary care empanelment. One fundamental element of empanelment is matching capacity with demand, which requires determining whether a primary care clinic or physician is "underempaneled" relative to a benchmark goal and therefore should accept new patients. It is difficult for an organization to achieve primary care physician (PCP) buy-in for regulation of panel size without a credible method of patient weighting to address physician concerns that raw counts do not accurately reflect panel variation. Weighted panel measurement may also assist health systems in prioritizing support staff to clinics and physicians with the highest work demand.

In summary, we have reported a novel clustering approach for primary care patients using routinely collected EHR data that can be used to create weighted panel sizes and dynamically load-balance PCPs and clinics. Our use of physician review of the clusters and predictive modeling suggests the algorithm identifies clinically meaningful phenotypes that are correlated with future primary care utilization. As health delivery and payment models shift to emphasize a population health orientation, weighting of primary care panels will assume greater importance for aligning primary care capacity and resources with variation in primary care work effort needed to care for different types of patients. Our weighting method attempts to capture this variation across patients in primary care work demand and may be implemented into population health analytic processes in a manner that allows near real-time calculation of weights in response to dynamic changes in patients' clinical activity.

Conflicts of Interest

Alvin Rajkomar reports serving as a Research Advisor to Google.

Multimedia Appendix 1

The absolute and change of within-group sum of squares with each additional cluster. The change in within-group sum of squares starts to level off at 4 groups.

[PNG File, 8KB - medinform_v4i4e29_app1.png]

Multimedia Appendix 2

Patient Characteristics of Each Utilization Phenotype in the Training Set (n=24,324).

[PDF File (Adobe PDF File), 18KB - medinform_v4i4e29_app2.pdf]

Multimedia Appendix 3

Patient Characteristics of each Primary Care Work Cluster in the Training Set (n=24,324).

[PDF File (Adobe PDF File), 17KB - medinform_v4i4e29_app3.pdf]

Multimedia Appendix 4

Log-Linear model of primary care office visits (without telephone visits) based on demographic variables and baseline utilization phenotype.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 18KB - medinform_v4i4e29_app4.pdf]

Multimedia Appendix 5

The unweighted and weighted patient counts across the primary care work clusters.

[PNG File, 74KB - medinform_v4i4e29_app5.png]

References

1. Petterson SM, Liaw WR, Phillips RLJ, Rabin DL, Meyers DS, Bazemore AW. Projecting US primary care physician workforce needs: 2010-2025. Ann Fam Med 2012;10(6):503-509 [FREE Full text] [doi: 10.1370/afm.1431] [Medline: 23149526]
2. Bodenheimer TS, Smith MD. Primary care: proposed solutions to the physician shortage without training more physicians. Health Aff (Millwood) 2013 Nov;32(11):1881-1886. [doi: 10.1377/hlthaff.2013.0234] [Medline: 24191075]
3. Altschuler J, Margolius D, Bodenheimer T, Grumbach K. Estimating a reasonable patient panel size for primary care physicians with team-based task delegation. Ann Fam Med 2012;10(5):396-400 [FREE Full text] [doi: 10.1370/afm.1400] [Medline: 22966102]
4. Bodenheimer T, Ghorob A, Willard-Grace R, Grumbach K. The 10 building blocks of high-performing primary care. Ann Fam Med 2014 Mar;12(2):166-171 [FREE Full text] [doi: 10.1370/afm.1616] [Medline: 24615313]
5. Grumbach K, Olayiwola JN. Patient empanelment: the importance of understanding who is at home in the medical home. J Am Board Fam Med 2015;28(2):170-172 [FREE Full text] [doi: 10.3122/jabfm.2015.02.150011] [Medline: 25748755]
6. Chung S, Eaton LJ, Luft HS. Standardizing primary care physician panels: is age and sex good enough? Am J Manag Care 2012 Jul;18(7):e262-e268. [Medline: 22823555]
7. Huntley AL, Johnson R, Purdy S, Valderas JM, Salisbury C. Measures of multimorbidity and morbidity burden for use in primary care and community settings: a systematic review and guide. Ann Fam Med 2012;10(2):134-141 [FREE Full text] [doi: 10.1370/afm.1363] [Medline: 22412005]
8. Ash AS, Ellis RP. Risk-adjusted payment and performance assessment for primary care. Medical Care 2012;50(8):643-653. [doi: 10.1097/MLR.0b013e3182549c74] [Medline: 22525609]
9. Ajorlou S, Shams I, Yang K. An analytics approach to designing patient centered medical homes. Health Care Manag Sci 2015 Mar;18(1):3-18. [doi: 10.1007/s10729-014-9287-x] [Medline: 24942633]
10. Pope GC, Kautter J, Ellis RP, Ash AS, Ayanian JZ, Lezzoni LI, et al. Risk adjustment of Medicare capitation payments using the CMS-HCC model. Health Care Financ Rev 2004;25(4):119-141 [FREE Full text] [Medline: 15493448]
11. Potts B, Adams R, Spadin M. Sustaining primary care practice: a model to calculate disease burden and adjust panel size. Perm J 2011;15(1):53-56 [FREE Full text] [Medline: 21505619]
12. Simborg DW. DRG creep: a new hospital-acquired disease. N Engl J Med 1981 Jun 25;304(26):1602-1604. [doi: 10.1056/NEJM198106253042611] [Medline: 7015136]

Abbreviations

AIC: Akaike information criterion
EHR: electronic health record
PCP: primary care physician
UCSF Health: University of California, San Francisco health system

Edited by G Eysenbach; submitted 22.08.16; peer-reviewed by T Kaji, CH Li; comments to author 14.09.16; revised version received 19.09.16; accepted 20.09.16; published 14.10.16.

Please cite as:
Rajkomar A, Yim JWL, Grumbach K, Parekh A
Weighting Primary Care Patient Panel Size: A Novel Electronic Health Record-Derived Measure Using Machine Learning
JMIR Med Inform 2016;4(4):e29
URL: http://medinform.jmir.org/2016/4/e29/
doi:10.2196/medinform.6530
PMID:27742603

©Alvin Rajkomar, Joanne Wing Lan Yim, Kevin Grumbach, Ami Parekh. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 14.10.2016. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction

XSL•FO
RenderX

in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Predictive Model for Medical Events Based on Contextual Embedding of Temporal Sequences

Wael Farhan[1], MS; Zhimu Wang[1,2], BA; Yingxiang Huang[1], BA; Shuang Wang[1], PhD; Fei Wang[3], PhD; Xiaoqian Jiang[1], PhD

[1]Health Sciences, Department of Biomedical Informatics, University of California - San Diego, La Jolla, CA, United States

[2]Department of Economics, Boston University, Boston, MA, United States

[3]Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, United States

Corresponding Author:
Wael Farhan, MS
Health Sciences
Department of Biomedical Informatics
University of California, San Diego
9500 Gilman Drive
La Jolla, CA, 92093
United States
Phone: 1 858 822 4931
Fax: 1 858 822 4931
Email: wyfarhan@gmail.com

Abstract

Background: Medical concepts are inherently ambiguous and error-prone due to human fallibility, which makes it hard for them to be fully used by classical machine learning methods (eg, for tasks like early stage disease prediction).

Objective: Our work was to create a new machine-friendly representation that resembles the semantics of medical concepts. We then developed a sequential predictive model for medical events based on this new representation.

Methods: We developed novel contextual embedding techniques to combine different medical events (eg, diagnoses, prescriptions, and labs tests). Each medical event is converted into a numerical vector that resembles its “semantics,” via which the similarity between medical events can be easily measured. We developed simple and effective predictive models based on these vectors to predict novel diagnoses.

Results: We evaluated our sequential prediction model (and standard learning methods) in estimating the risk of potential diseases based on our contextual embedding representation. Our model achieved an area under the receiver operating characteristic (ROC) curve (AUC) of 0.79 on chronic systolic heart failure and an average AUC of 0.67 (over the 80 most common diagnoses) using the Medical Information Mart for Intensive Care III (MIMIC-III) dataset.

Conclusions: We propose a general early prognosis predictor for 80 different diagnoses. Our method computes numeric representation for each medical event to uncover the potential meaning of those events. Our results demonstrate the efficiency of the proposed method, which will benefit patients and physicians by offering more accurate diagnosis.

(JMIR Med Inform 2016;4(4):e39) doi:10.2196/medinform.5977

KEYWORDS

clinical decision support; early classification; temporal phenotyping; clinical event context embedding

Introduction

Background

The large collection of healthcare data has brought tremendous opportunities and challenges to health care research [1]. The goal is to prevent and treat diseases by taking into account individual variabilities, which include genome, environment, and lifestyle [2]. There are many difficulties in making use of a large amount of health care data from heterogeneous sources with different characteristics (high dimensional, temporal, sparse, irregular, etc). The traditional data analysis methods (often developed for clean and well-structured data) do not fit these challenges well and may not be able to effectively explore

XSL•FO
RenderX

the rich information in the massive health care data. Most of the existing models treat different medical events as distinct symbols without considering their correlations, and therefore are limited in terms of representation power [3-7]. For example, it is hard for those methods to use the correlation among different types of events (eg, the similarity between a prescription and a diagnosis, or an abnormal lab and a diagnosis). Indeed, many models assume a vector-based representation for every patient, where each dimension corresponds to a specific medical event. Such representation loses the temporal context information for each medical event, which could be informative for impending disease conditions.

Diagnoses share common symptoms making them enigmatic and hard to differentiate. Physicians might have a hard time discovering potential risks. Recent studies show that most diagnostic errors have been associated with flaws in clinical reasoning and empirically prove the evidence between cognitive factors and diagnostic mistakes [8,9]. In 25% of the records of patients with a high-risk diagnosis, high-information clinical findings were present before the high-risk diagnosis was established [10]. Our predictive model aims to counterbalance cognitive biases by suggesting possible diagnoses based on the patient's medical history. We combine data from different sources in an innovative way, which synthesize information more comprehensively than existing models. Our model is more accurate than most predictive models in the literature and it is less computationally expensive.

With the above considerations, we introduced a new representation for electronic health records (EHR) that was context-aware and combines heterogeneous medical events in a uniform space. Here, the "context" was defined with respect to each medical event in the patient EHR. The context around an event *A* is the order of medical events happening before and after *A* within the patient EHR corpus. For each patient, through the concatenation of all medical events in his or her EHR according to their sequential timestamps (without considering the order of tied events), we obtained a "timeline" describing all historical conditions of the patient. While generating context, we lost the exact time at which each event occurred. Therefore, the context around a specific medical event in the timeline was similar to the context around a word in a narrative text.

How to derive effective word representations by incorporating contextual information is a fundamental problem in natural language processing and has been extensively studied [11-13]. One recent advance is the "Word2Vec" technique that trains a 2-layer neural network from a text corpus to map each word into a vector space encoding the word's contextual correlations [14,15]. The similarities (usually computed by the cosine distance over the embedded vector space) reflect the contextual associations (eg, words *A* and *B* with high similarity suggest that they tend to appear in the same context). Word2Vec is able to extract event semantics despite the relatively small training corpus. We extended Word2Vec to support dynamic windows to handle the temporal nature of medical events.

Based on the contextual embedding representation, we developed 3 models to predict the 80 most common diagnoses based on Medical Information Mart for Intensive Care III (MIMIC-III) dataset. The goal of this study was to predict the onset risk of each diagnosis based on historical patient records. Our model achieves an area under the receiver operating curve (ROC) curve (AUC) higher than 0.65 for half of the 80 diagnoses. We further introduced time decay factors in the model to reflect the fact that more recent events have a bigger impact on the prediction. Our model was also able to learn bioequivalent drugs (and medical events) and build the semantic relationship, which cannot be fulfilled with most existing predictive models.

In this paper, we encountered a more challenging task than previous work mentioned in the next section. Here, we built a novel diagnosis predictor, which means our model was predicting diagnoses that do not occur in patient history. Most of chronic disease will eventually be listed on every admission for that patient, predicting the same diagnosis again will enhance the performance of our predictor but will not add anything new for the physician treating that patient. Nonetheless, we ran predictor against all diagnoses (ie, not restricted to novel ones) to be able to compare it with previous work. We achieved a mean AUC of 0.76 for 80 diagnoses.

Previous Work

A substantial amount of work has been conducted on systems to support clinical decisions using predictive models. For example, Gottlieb et al [3] proposed a method for inferring medical diagnoses from patient similarities using patient history, blood tests, electrocardiography, age, and gender information. However, their method can only predict discharge codes at international classification of diseases (ICD)-9 level 1, which are relatively generic and cannot differentiate among a wide range of diverse diagnoses. In risk prediction with EHR, Cheng et al [16] used convolutional neural network with a temporal fusion to predict congestive heart failure and chronic obstructive pulmonary disease within the next 180 days. Their approach can only handle 2 diagnoses and achieved an AUC of less than 0.77. Ghalwash et al [17] extracted multivariate interpretable patterns for early diagnosis. They constructed key shapeletes (a time series subsequence) to represent each class of early classification using an optimization-based approach. This technique is computationally expensive and would not work efficiently with a large dataset, therefore, they only focused on a small number of diagnoses. By taking advantage of a different set of inputs, functional magnetic resonance imaging (fMRI) images, Wang et al proposed high-order sparse logistic regression and multilinear sparse logistic regression [18,19] for early detection of Alzheimer disease and congestive heart failure. Their results surpassed standard learning algorithms, such as nearest neighbor, support vector machines (SVM), logistic regression (LR), and sparse logistic regression. But not all patients have fMRI images within EHR, thus their models are only limited to a small subset of patients. Taslimitehrani et al [20] constructed a logistic regression model using CPXR(log) method (short for contrast pattern aided logistic regression) to predict mortality rate in heart failure patient. They consulted a cardiologist and a cardiovascular epidemiologist to identify patient cohort from EHR data collected from patients admitted to the Mayo Clinic between 1993 and 2013. Their model is specific and can only be extended to different diagnoses after

XSL•FO
RenderX

consulting specialists. Recently, Lipton et al [21] used long short-term memory (LSTM) recurrent neural network for a multilabel classification of diagnosis in the pediatric intensive care unit, which demonstrated improved performance over a set of standard learning methods. They trained LSTM neural network (ie, a special recurrent neural network, which has a forget gate to capture long-term dependency) on variable length inputs of large size. Nevertheless, their model is a black box, which cannot be interpreted by human experts.

There is also some related work on feature representation. Tran et al [22] presented a generative model based on nonnegative Restricted Boltzmann Machine to learn low-dimensional representations of the medical events from electronic medical records (EMRs). Their model assumes EMRs are aggregated into regular time intervals and captures the global temporal dependency structures of the events. Another work by Che et al [23] explored deep learning applications to the problem of discovery and detection of characteristic patterns of physiology in clinical time series. They applied deep feed-forward neural network with fully connected layers using graph Laplacian priors and developed an efficient incremental training procedure to detect physiological patterns of increasing length, which demonstrated good AUCs. Using a similar approach, Liu et al [24] extracted temporal phenotypes from longitudinal EHR using a graph-based framework. They represented each patient’s history using a temporal graph, where each node serves as a medical event and edges are constructed based on the temporal order of events. Using those temporal graphs, they identified the most significant and interpretable subgraph basis as phenotypes, which is used later as a feature set for their predictive model. But their method has only been applied to a small cohort associated with congestive heart failure.

The context-aware representation proposed in this paper provides a new way of combining data and building predictive models. We developed several methods on top of the novel representation and achieved a high AUC. As mentioned earlier, none of the previous work tackled the challenge of predicting a novel diagnosis. In this paper, we show that our model is able to predict a diagnosis that was not previously identified. Also, our model is highly generalizable, which can predict multiple diseases without having to tune parameters for each one of them.

Methods

Temporal Sequence Construction

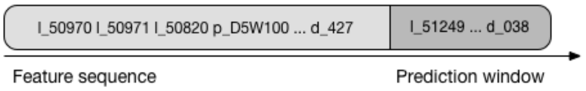
In this section, we will present the proposed sequential prediction framework by starting with explanation about what the components of a sequence are and how the sequential prediction is formulated.

In our model, a sequence was defined as a combination of lab tests, prescriptions, and diagnoses that were performed, ordered, or assigned to a patient in multiple hospital admissions. Lab tests and prescriptions were represented by unique identifiers defined by the dataset. But because two tied events could have the same identifier we added ‘ *l* _’ at the beginning of lab tests key and ‘ *p* _’ for prescriptions. Diagnoses, on the other hand, were all represented with their ICD-9 code prefixed with ‘ *d* _’. To conserve part of the temporal information, we sorted those events from oldest to latest. Hence, we lost the exact timestamp at which the event happened. A patient sequence contained data from multiple admissions that happened within a year apart from each other. We sliced the most recent admission out of the sequence and used its diagnoses as gold standard in the prediction phase, while preceding admission events are used as features. A graphical illustration of a sequence is depicted in Figure 1.

Unlike earlier work, in this paper we did not preprocess diagnosis ICD-9 level to generalize them at one level. Instead, we kept the ICD exactly as identified by the physician. For example, “pneumonia” (486) is a level 3 diagnosis and “anemia in chronic kidney disease” (285.21) is a level 5; all were kept as unique events in the same sequences. This way, our predictor will identify the diagnosis in the same specificity level as diagnosed by the physicians.

Also, due to the nature of medicine, some medical events are extremely rare in the population. Hence, it would be hard to extract common patterns from a very small sample. For our experiments, we excluded events that appear in less than 1% of the total number of sequences.

Figure 1. Sequence construction.



Contextual Embedding

Word2Vec [15], a tool created to learn word embeddings from a large corpus of text, has recently gained popularity. It has mainly been applied in natural language processing to generate continuous vector representation for each word. The distances between these words (in the vector space) describe the similarities of those words. A well-known example of the so-called “semantic relationship” presented in the original paper is that queen to king has almost same distance like woman to man [15]. Another popular semantic relationship learned using the same model is reported as “V[France] – V[Paris] ≈ V[Germany] – V[Berlin]” [8], where *V* is the vector representation of the word.

Word2Vec, in its core, depends on 2 parameters: size and window; size defines the dimensionality of the vector representation, while window is the maximum distance between a word and its predicate word in one sentence. Word2Vec supports 2 modes of operation [25]: (1) Continuous Bag of Words: the input to the model is a collection of words, and the model would predict the missing word, and therefore, it can predict a word given its context as illustrated in Figure 2 a; and

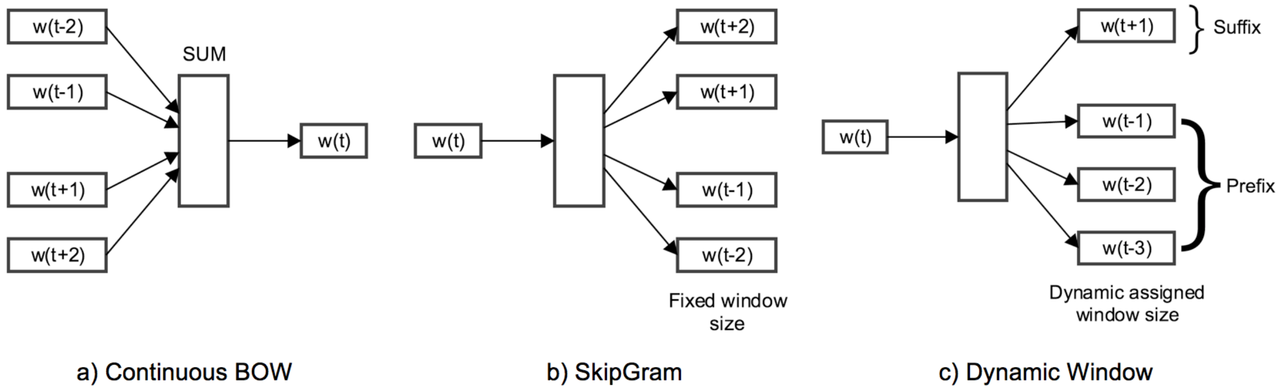
XSL•FO
RenderX

(2) Skip-Gram: the target word is now in the input to the model, and the context words are going to be predicted, as illustrated in Figure 2 b.

In the proposed model, we extend Word2Vec to support one extra mode as follows: Dynamic Window: a customized mode in our experiment defines different windows for words in the sequence as prefix (preceding words) and suffix (succeeding words) as illustrated in Figure 2 c.

In our paper, we used Word2Vec to generate vector representation for each medical event by feeding it with the medical event sequences discussed in the previous section. With Word2Vec technique, we can extract event semantics from a relatively small corpus.

Figure 2. Different Word2Vec modes. (a) and (b) are the Continuous Bag of Words (CBOW) and SkipGram modes, which have been widely used in neurolinguistic programing (NLP) problems; (c) a new and more flexible mode to support models using dynamic window.



Learning Methods

We present the proposed predictive methods in this section. For each method, we used the training set to learn binary classification models for diagnoses of interest. Those binary classification models calculate the probability of having a future diagnosis given test sequences. A test sequence will end up with multiple predictions, one for each diagnosis. Each diagnosis prediction is completely independent from other diagnoses, formulating our approach as multiclass classification problem. All learning methods make use of the contextual representation generated by Word2Vec. We passed patient sequences from the training set into Word2Vec to learn a contextual vector representation for each medical event.

Collaborative Filtering

In this method, we leveraged a recommendation system [26] that calculates patient-patient projection similarity. Each patient record in a training set was projected into the vector space by summing up event vectors in its sequence multiplied by the temporal factor. Intuitively, patients with similar history projections are more likely to foretell the future more than others. This information was used in the decision of what diagnosis a patient might get.

For prediction, we projected the test patient sequence exactly like training records. Then, we found the patients with the most similar projections. We calculated the probability based on weighted voting, where the weight is the cosine similarity of the 2 patients (Figure 3).

Where *s* is a patient sequence, *d* is a diagnosis, p_d corresponds to all patients in training set who end up with diagnosis *d*, and *p* corresponds to all patients.

Figure 3. Collaborative filtering weighted voting.

$$\hat{y}(s,d) = \frac{\sum_{p_d} max[0, cosine(s,p_d)]}{\sum_{p} max[0, cosine(s,p)]},$$

Patient-Diagnosis Event Similarity

In the patient-diagnosis event similarity (PDES) prediction method, we used the generated vector representation to build *S*, a cosine similarity matrix. *S* is a (*N*×*D*) matrix, where *N* is the number of all medical events and *D* is the number of diagnoses. For example, *S['d_428','l_50862']* is the cosine similarity between heart failure and albumin blood test.

To predict the diagnosis given in a patient sequence, we first generated patient event vector of length N by simply summing one-hot representation (eg, mapping the medical events to vectors of length N, where the nth digit is an indicator of that medical event) of its events multiplied by temporal factor, to emphasize recent events. Then, we use this array to find the similarity of that patient with a particular diagnosis using the equation in Figure 4.

Where *s* is a patient sequence, *d* is a diagnosis, σ is a normalization constant, v_d is a column in the similarity matrix corresponding to the diagnosis *d*, *c* is a medical event, λ is the decay factor and t_c is time passed from the latest event. is the one-hot vector representation of *c*. The term $e^{-\lambda tc}$ is used to

XSL•FO
RenderX

account for the decay of impact of medical histories like in the previous example. Figure 5 depicts the prediction methodology of PDES.

The higher the similarity, the more likely a patient will get the diagnosis in the next visit. It is possible to get negative similarity values, but empirical evaluation showed that converting negative similarities to zero achieved better performance. There are very few hyperparameters that need tuning: Word2Vec size and window parameters, and λ, the decay factor.

One of the issues with this approach is that it does not take full advantage of the semantic similarity. Consider 2 similar prescriptions, the cosine similarity of them with respect to a particular diagnosis will be almost the same. If a patient happens to be treated with the first prescription and not the second, then, the patient representation will have a value of zero at the one-hot representation of the second prescription. Hence, the result of the dot product in Figure 6 will falsely diminish, reducing the probability of that diagnosis. We will overcome this problem in the next method.

Figure 4. Patient-diagnosis event similarity.

$$\hat{y}(s,d) = \tfrac{1}{\sigma}(v_d^T \sum_{c \in s} e^{-\lambda t_c} \vec{0}_{-c}),$$

Figure 5. Patient diagnosis event similarity.

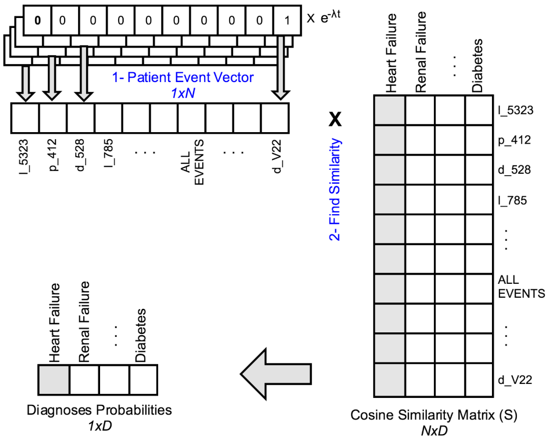
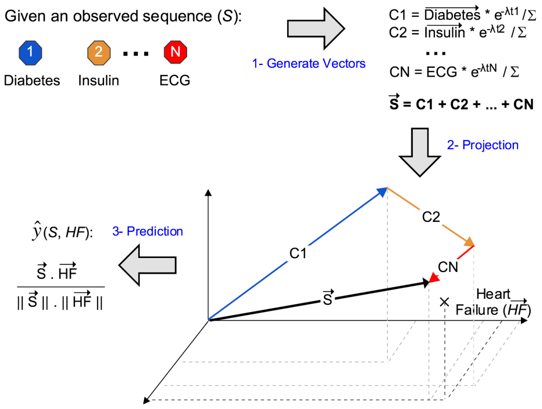


Figure 6. Patient-diagnosis projection similarity, where Σ is the summation of temporal factors.



Patient-Diagnosis Projection Similarity

Based on the vector representation, we also proposed another prediction method, patient-diagnosis projection similarity (PDPS), where we project patient sequence into the vector space, bearing in mind the temporal impact. Then, we computed the cosine similarity between the patient vector and the diagnosis vector. The equation in Figure 7 demonstrates the prediction method.

Where the is the vector contextual representation of diagnosis *d* in the vector space, and is the vector contextual representation of a medical event in patient sequence. Figure 6 illustrates the prediction methodology used in PDPS similarity method. PDPS can solve the problem of nonidentical similar events faced by PDES. Here, patient projection is unaffected by similar events; whether the patient got the first prescription or the second, PDPS would still add an equivalent vector into the patient projection.

XSL•FO
RenderX

Figure 7. Patient-diagnosis projection similarity.

$$\hat{y}(s,d) = cosine\_similarity(\vec{V}_d\ ,\ [\frac{\sum_{c\in s} \vec{V}_c e^{-\lambda t_c}}{\sum_{c\in s} e^{-\lambda t_c}}]),$$

Results

Data (Medical Information Mart for Intensive Care III)

In this section, we evaluate the proposed methods with a real-world dataset. We present the results of these experiments and discuss the choice of hyperparameters. We also compare the results of different models, diagnoses, and datasets. We compare our results with standard learning methods/algorithms that do not make use of the contextual representation. We will begin this section by introducing the dataset we used.

To test the proposed methods, we explored MIMIC-III database [27], which contains health-related data associated with 46,520 patients and 58,976 admissions to the intensive care unit of Beth Israel Deaconess Medical Center between 2001 and 2012. The database includes detailed information about patients, such as demographics, admissions, lab test results, prescription records, procedures, and discharge ICD-9 diagnoses.

Because we wanted to predict the next diagnosis, we excluded the patients who were only admitted once. We also eliminated rare lab tests and prescriptions that only happened in less than 50 admissions. In total, we select 204 most common lab events that flagged as abnormal, 1338 most common prescriptions, 826 most common diagnoses, 274 most common conditions, and 171 most common symptoms. After applying the method introduced in the Temporal Sequence Construction section, we constructed 5642 temporal sequences using medical records of 5195 patients. The total number of sequences was larger than the number of patients because we used a threshold of 1 year as the medical history cut-off. Hence, a patient could have multiple sequences if admissions happen more than 1 year apart.

Baselines

As a sanity check, we needed a proof to make sure that our models were more beneficial than common learning models. So we decided to compare our results with baseline models. First, we converted medical events into one-hot representation vectors. Then, we generated patient vectors by summing up the one-hot representation vectors of its events. These vectors served as input features, while the label was the binary value that indicated whether a diagnosis was found in the last admission. We generated one label vector for each diagnosis and ran our learning algorithm once for each diagnosis.

We explored multiple baseline models by passing our features through SVM, LR, and decision trees learning models. We also applied decay factor just as described in PDES model. LR with decay was able to achieve the highest results. Hence, we decided to adopt it as our baseline.

Performance

We applied the 4 methods to the MIMIC dataset. We adopted AUC, accuracy, and *F*-score as measurements to compare different models. We used 10-fold cross-validation to evaluate each model.

Other than the baseline model LR, all models we proposed incorporated the vector representations of medical events from Word2Vec. For visualization purposes, we limited the dimensions of the hyperspace to 2 dimensions. Figure 8 illustrates the limited contextual representation color coded by event type. Vector representations constructed by Word2Vec were able to capture semantic meaning of medical events. Word2Vec clusters events based on their type as shown in the figure. In addition, it was able to capture closely similar events, for example, the cosine similarity of *’p_WARF2’* (Warfarin 2-mg Tab) and *’p_WARF1’* (Warfarin 1-mg tab) was 0.924. All prescriptions starting with *’p_WARF’* were close to each other around the point (0.5, 0.2). This representation simplifies learning because it groups similar events by unified content.

Our experiments included predicting the 80 most common diagnoses for each patient. More formally, we constructed a multilabel classification problem where each patient sequence could be labeled with multiple diagnoses. A patient is labeled with a diagnosis if and only if that particular diagnosis happens in the final admission (ie, prediction window). We selected 4 diagnoses to discuss in the paper, which are displayed in Table 1 with AUC for each diagnosis in each model. From the table, it is noticed that PDPS achieves the highest performance in most cases. The full results can be found in Multimedia Appendix 1. Figure 9 contains 7 selected ROC curves collected from the entire 80 diagnoses. This figure shows how our learning method performs differently on various diagnoses.

XSL•FO
RenderX

Figure 8. Medical event contextual representation displayed in 2 dimensions for visualization purposes. These dimensions are arbitrary learned by Word2Vec.

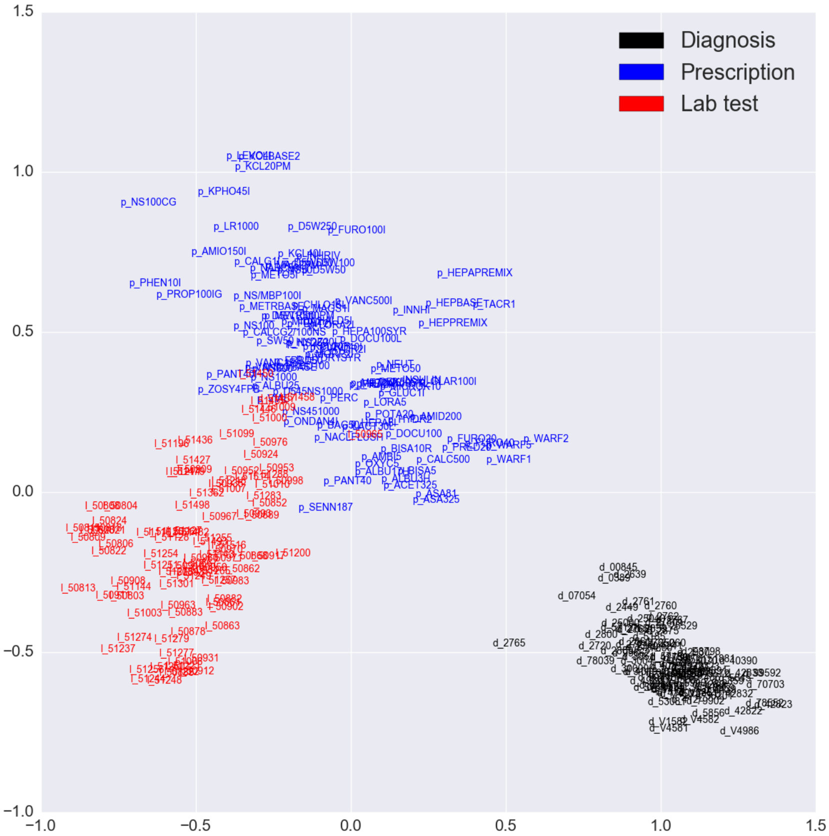


Table 1. Sample results of MIMIC-III.

	Chronic systolic heart failure (485.22)			Acute posthemorrhagic anemia (285.1)			Hyperlipidemia not elsewhere classifiable/not otherwise specified (272.4)			Septicemia not otherwise specified (038.9)		
	AUC[a]	ACC[b]	*F*-score	AUC	ACC	*F*-score	AUC	ACC	*F*-score	AUC	ACC	*F*-score
LR[c]	0.780	0.947	0.237[d]	0.550	0.779	0.159	0.699	0.737	0.224	0.593	0.651	0.217
CF[e]	0.784	0.849	0.145	0.581	0.815[d]	0.152	0.733[d]	0.772	0.254[d]	0.641	0.632	0.225
PDES[f]	0.793	0.869	0.158	0.579	0.408	0.153	0.702	0.763	0.221	0.648	0.682	0.239
PDPS[g]	0.795[d]	0.953[d]	0.213	0.618[d]	0.786	0.175[d]	0.723	0.851[d]	0.238	0.652[d]	0.720[d]	0.242[d]

[a] AUC: area under the receiver operating characteristic curve.

[b] ACC: accuracy.

[c] LR: logistical regression.

[d] The highest value between the four different methods.

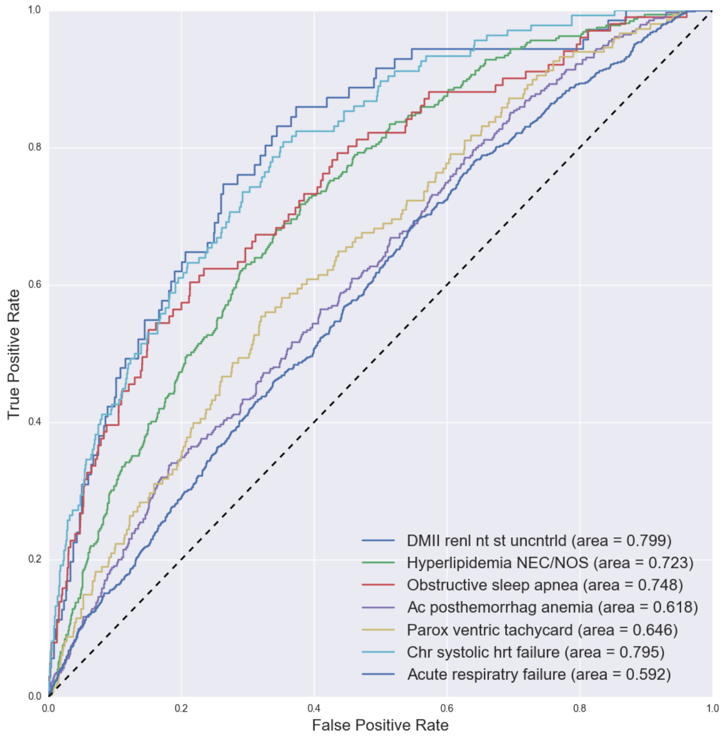
[e] CF: collaborative filtering.

[f] PDES: patient-diagnosis event similarity.

[g] PDPS: patient-diagnosis projection similarity.

XSL•FO
RenderX

Figure 9. Patient–diagnosis projection similarity (PDPS) receiver operating characteristic (ROC) curves and their corresponding area under the receiver operating characteristic curve (AUCs) for each disease prediction.



The outcome of each binary diagnosis predictor was a probability between 0 and 1. We computed a distinct threshold for each diagnosis, above which a patient was labeled as positive. The threshold was calculated such that it optimizes the F1 score (ie, Youden index [28]). Finally, the accuracy gets computed after labeling test patients. Figure 10 displays AUC results of 30 different diagnoses using PDPS. As can be seen from the graph, our results are robust across diagnoses and models, and demonstrated clear performance advantage over other methods in comparison.

We investigated how our predictor works by analyzing the true positive sequences of patients to find a medical justification behind each diagnosis. For each diagnosis, we computed the top medical events that our predictor used as the leading cause. Most findings were precise and clinically insightful (thanks to our medical doctor collaborators for examination). We list a few examples here. Chronic kidney disease (CKD) is predicted after finding late manifestations of joint, soft tissue, and bone problems coexist (musculoskeletal). In addition, over the counter pain killers (nonsteriodal ant-inflammatory drugs), can cause CKD; however, this problem often goes unrecognized by health care providers, especially when they do not check kidney function. Another example is pneumonia, where our predictor associates glossitis, which can lead to problems in protecting patients' airways, with pneumonia. Chest deformity can damage blood vessels (capillaries) in the lungs, allowing more fluid to pass into the lungs, making the patient more sensitive to bacteria, viruses, fungi, or parasites infections. Vocal cord diseases can also lead to pneumonia so as autosomal anomalies where abnormal chromosomes make patients at increased susceptibility to respiratory disease like pneumonia and other infectious disease. Another example, obstructive sleep apnea is predicted through structural and mechanical problems like acute tracheitis without mention of obstruction, scoliosis, and obesity, in addition to inflammation in the nasal membranes like allergic rhinitis and poisoning by opiates and related narcotics, which cause sleep disturbance and hypoventilation (decrease in respiratory rate).

Yet another example is that cirrhosis of the liver without ETOH linked with several hepatitis C disorders. Hepatitis C can be a precursor to nonalcoholic cirrhosis. Malignancy of the rectosigmoid junction would rarely cause cirrhosis, but can sometimes result in liver metastasis that can cause laboratory abnormalities similar to those found in cirrhosis–that is why our predictor slightly linked them together. Our predictor was able to learn patterns of these diagnoses without the supervision of a medical practitioner.

Decay - Temporal Effect

Figure 11 illustrates the effect of adding temporal factor to the PDPS prediction model. Adding temporal factor forced the model to focus more on the recent events and to leave older ones with less influence. The main observation here is that different diseases behave distinctly. Some diagnoses like "volume depletion disorder" and "anemia" decreased in AUC as we increased decay factor, which means that those diseases are predicted more accurately by looking at the entire patient history. Others like "end-stage renal disease" increased AUC when increasing the decay factor, which implies that the model had to focus on the last few events to be able to predict it. Most of the diagnoses like "aortic valve disorder" and "hyperlipidemia" had a bell shaped curve with different optimal decay value. This phenomenon applies to all methods including the baseline.

XSL•FO
RenderX

Figure 10. Patient-diagnosis projection similarity (PDPS) area under the receiver operating characteristic curve (AUC) of 30 diagnoses on the medical information mart for intensive care III (MIMIC III) dataset. (A) septicemia NOS, (B) hypothyroidism not otherwise specified (NOS), (C) protein-cal malnutr NOS, (D) pure hypercholesterolem, (E) hyperlipidemia not elsewhere classifiable (NEC)/NOS, (F) hyposmolality, (G) acidosis, (H) Ac posthemorrhag anemia, (I) anemia-other chronic dis, (J) thrombocytopenia NOS, (K) depressive disorder NEC, (L) obstructive sleep apnea, (M) hypertension NOS, (N) Hy kid NOS w cr kid I-IV, (O) Hyp kid NOS w cr kid V, (P) old myocardial infarct, (Q) Crnry athrscl natve vssl, (R) atrial fibrillation, (S) congestive heart failure NOS, (T) pneumonia, organism NOS, (U) Chr airway obstruct NEC, (V) food/vomit pneumonitis, (W) pleural effusion NOS, (X) pulmonary collapse, (Y) cirrhosis of liver NOS, (Z) acute kidney failure NOS, (a) end-stage renal disease, (b) chronic kidney dis NOS, (c) osteoporosis NOS, and (d) Surg compl-heart.

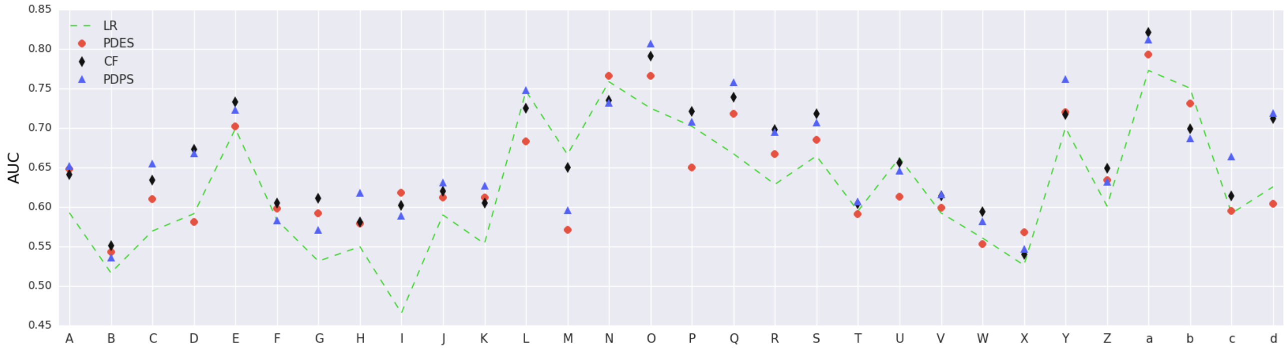
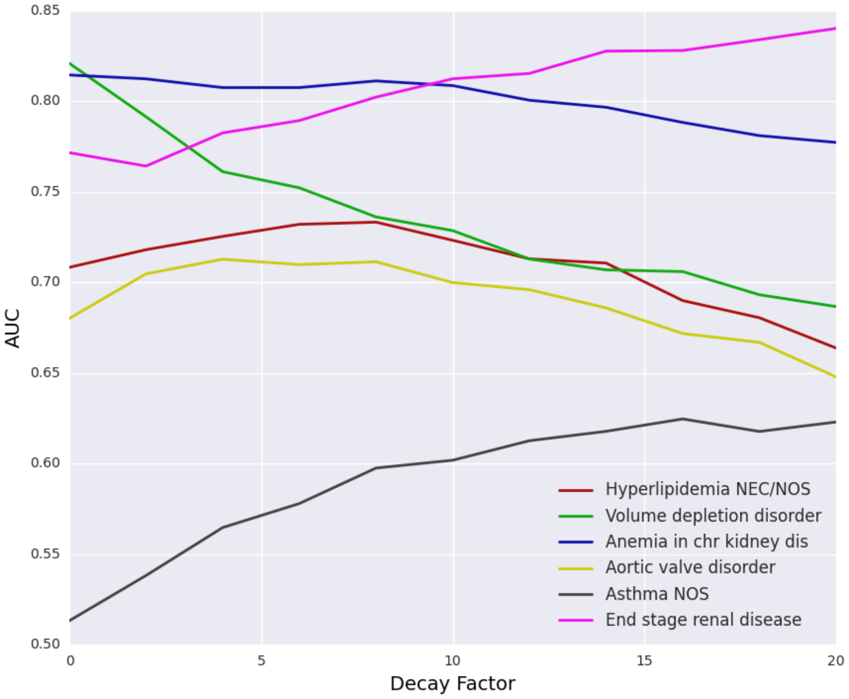


Figure 11. Effect of decay on patient-diagnosis projection similarity (PDPS) similarity.



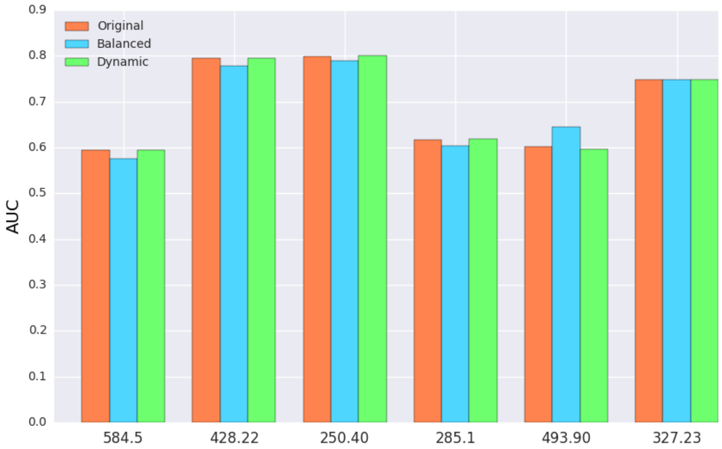
Data Balancing

Health data are often uneven where some diagnoses are more common than others. For example, in the MIMIC-III dataset, “gout” is less common than “congestive heart failure.” This can affect the downstream predictive models and we tried to mitigate it by balancing the dataset. However, balancing the dataset was not an easy task because admissions tended to be labeled with multiple diagnoses, for example, diabetes (ICD-9: 250.00) and congestive heart failure (ICD-9: 428.0). Therefore, when we try to balance an infrequent diagnosis, by duplicating some of its sequences randomly, we increase the rate of other diagnoses that happened with the infrequent one. We approximated the balance by making sure that each diagnosis appeared in at least 8% of the total sequences. This step was done by duplicating random samples that contained infrequent diagnoses until all diagnoses passed the 8% threshold.

As shown in Figure 12, balancing had small impact on the overall performance. Context representation did not change a lot from adding the same sequence again, and that explains why our model did not benefit from rectifying skewness.

XSL•FO
RenderX

Figure 12. Effect of balancing the dataset on patient-diagnosis projection similarity (PDPS). 584.5 Ac kidny fail, tubr necr, 428.22 Chr systolic hrt failure, 250.40 DMII renl nt st uncntrld, 285.1 Ac posthemorrhag anemia, 493.90 Asthma NOS, 327.23 Obstructive sleep apnea.



Dynamic Window

Recall that dynamic window defines different window sizes for each word in the sequence. In PDPS, we defined the window to be 365 days, so any medical event that happened before that would be discarded, so that they have no influence on the contextual representation. As can be seen from Figure 12, there is a minor impact on overall performance because we believed that the dynamic window was being overshadowed by the temporal decay. In other words, the influence of old events was limited due to our adaptation of the temporal factor, eliminating it by dynamic window was not going to bring a significant change.

The results show that a predictive models using semantic extraction worked better than baseline learning methods. The PDPS method achieved the highest mean performance across 80 different diagnoses. Each diagnosis reached its highest AUC on a different decay constant lambda, this variation depended on the nature of the disease. We also exposed different variations that included dynamic window and balancing the dataset.

Discussion

Limitations and Future Work

The proposed studies have several limitations. When making predictions for our datasets, we neglected demographic information such as age, gender, and race. One way of incorporating this information is by injecting extra words in the sequences, for example, gender could be represented as *'g_Male'* and *'g_Female'*. We believe that some demographic information is already embedded within the medical event vector representation, for example, normal delivery (with ICD-9 code: 650) would also imply that the patient is a female. Therefore, adding vocabulary to explicitly identify the demographics may not improve the model significantly. We will test this hypothesis in future work.

Most learning models deal with a group of hyperparameters like decay factor, window, size, and space dimension. Tuning those parameters consumes a considerable amount of time and effort, especially for collaborative filtering. PDES and PDPS are substantially faster so we are able to tune the parameters and reuse them for collaborative filtering method.

Medical error is one of the issues with which all early prognosis predictors have to deal [26]. Medical error might include misdiagnosis, delayed, inaccurate, or incomplete diagnosis. Diseases related to inflammation, autoimmune, or mild infection (with ICD-9 codes: 424, 507, 511, etc) has no specific symptoms; need extensive lab work; and could still be incorrectly analyzed. When training contextual representation, a sequence in the training set with misdiagnosis could slightly modify vector projection of medical events, which might be negligible. On the other hand, a misdiagnosed test sequence could alter the overall performance. There are some diseases, such as pneumonia (486) and septicemia (038), which develop quickly and do not have a history pattern. Thus PDPS does not do very well (AUC slightly over 0.60 for those difficult cases). We might need to develop new and customized models to predict these special cases.

Another limitation of our approach is that it assumes the sequence events are sampled at the same frequency (without considering the order of tied events), which means the temporal effect is not accurately represented. We can solve this problem by incorporating each event with timestamp in combination with dynamic window for the accurate representation.

Conclusion

We developed a sequential prediction model of clinical phenotypes based on contextual embeddings of medical events. Using the vector representation as features for our PDPS model, we were able to achieve a mean AUC of 0.67 and a median AUC of 0.65 (AUC ranging between 0.54 and 0.85) on 80 diagnoses from MIMIC dataset. The results demonstrated that learning EHR could benefit from abstracted contextual embeddings, which also preserved the semantics for human interpretation.

Our approach suggested a new way to learn EHR using contextual embedding methods, where we believe there is still much to discover. In this paper, we explored a set of prediction methods that exploit medical event embeddings. The experimental results showed that our best predictor is able to

XSL•FO
RenderX

efficiently learn 14,080 medical cases with 10-fold cross validation under 15 minutes as well as achieved an AUC better than most state-of-the-art methods. We recognize that some diagnoses are still hard to predict either due to their medical complexity and wind up misdiagnosed or due to their sudden unexpected nature. In future work, we plan to focus on making temporal factors more accurate and fusing demographic information within patient medical event sequences.

Acknowledgments

SW and XJ were partially supported by National Human Genome Research Institute Grants R00HG008175 and R01HG007078, the National Institute of General Medical Sciences R01GM114612, National Library of Medicine Grants R21LM012060, and U54HL108460; YW was supported by T15LM011271. FW is partially supported by National Science Foundation under Grant Number IIS-1650723. We thank Rebecca Marmor, MD and Shatha Farhan, MD for providing clinical insights.

Conflicts of Interest

None declared.

Multimedia Appendix 1

MIMIC III Results for All 80 Diagnoses.

[PDF File (Adobe PDF File), 44KB - medinform_v4i4e39_app1.pdf]

References

1. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015;372:793-795. [doi: 10.1056/NEJMp1500523] [Medline: 25635347]
2. National Institutes of Health (NIH). Precision Medicine Initiative URL: https://www.nih.gov/precision-medicine-initiative-cohort-program [accessed 2016-11-09] [WebCite Cache ID 6lthLxvsx]
3. Gottlieb A, Stein GY, Ruppin E, Altman RB, Sharan R. A method for inferring medical diagnoses from patient similarities. BMC Med 2013;11:194 [FREE Full text] [doi: 10.1186/1741-7015-11-194] [Medline: 24004670]
4. Jiang X, Boxwala AA, El-Kareh R, Kim J, Ohno-Machado L. A patient-driven adaptive prediction technique to improve personalized risk estimation for clinical decision support. J Am Med Inform Assoc 2012;19:e137-e144 [FREE Full text] [doi: 10.1136/amiajnl-2011-000751] [Medline: 22493049]
5. Alodadi M, Janeja P. Similarity in patient support forums using TF-IDF and cosine similarity metrics. : IEEE; 2015 Presented at: International Conference on Healthcare Informatics (ICHI); October 21-23, 2015; Dallas, Tx p. 521-522 URL: http://ieeexplore.ieee.org/document/7349760/
6. Sun J, Wang F, Hu J, Edabollahi S. Supervised patient similarity measure of heterogeneous patient records. SIGKDD Explor Newsl 2012;14:16-24 [FREE Full text] [doi: 10.1145/2408736.2408740]
7. Luukka P. Similarity classifier in diagnosis of bladder cancer. Comput Methods Programs Biomed 2008;89:43-49 [FREE Full text] [doi: 10.1016/j.cmpb.2007.10.001]
8. van den Berge K, Mamede S. Cognitive diagnostic error in internal medicine. Eur J Intern Med 2013;24:525-529 [FREE Full text] [doi: 10.1016/j.ejim.2013.03.006] [Medline: 23566942]
9. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Arch Intern Med 2005;165:1493-1499 [FREE Full text] [doi: 10.1001/archinte.165.13.1493] [Medline: 16009864]
10. Feldman MJ, Hoffer EP, Barnett GO, Kim RJ, Famiglietti KT, Chueh H. Presence of key findings in the medical record prior to a documented high-risk diagnosis. J Am Med Inform Assoc 2012;19:591-596 [FREE Full text] [doi: 10.1136/amiajnl-2011-000375] [Medline: 22431555]
11. Klein D, Manning CD. Natural language grammar induction with a generative constituent-context model. Pattern Recognition 2005;38:1407-1419 [FREE Full text] [doi: 10.1016/j.patcog.2004.03.023]
12. Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: an architecture for development of robust HLT applications. 2002 Presented at: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02; July 6-12, 2002; Philadelphia, PA p. 168-175 URL: http://dl.acm.org/citation.cfm?id=1073112 [doi: 10.3115/1073083.1073112]
13. Pustejovsky J, Boguraev B. Lexical knowledge representation and natural language processing. Artificial Intelligence 1993;63:193-223 [FREE Full text] [doi: 10.1016/0004-3702(93)90017-6]
14. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. 2013 Presented at: Advances in Neural Information Processing Systems 26 (NIPS 2013); December 5-10, 2013; Harrahs and Harveys, Lake Tahoe, NV p. 3111-3119 URL: http://papers.nips.cc/paper/5021-distributed-representations
15. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. Cornell University Library. 2013. URL: http://arxiv.org/abs/1301.3781 [accessed 2016-11-11] [WebCite Cache ID 6lvwqUWO3]

XSL•FO
RenderX

16. Cheng Y, Wang F, Zhang P, Hu J. Risk prediction with electronic health records: a deep learning approach. 2016 Presented at: SIAM International Conference on Data Mining; May 5-7, 2016; Miami, FL URL: http://astro.temple.edu/~tua87106/sdm16.pdf
17. Ghalwash MF, Radosavljevic V, Obradovic Z. Extraction of interpretable multivariate patterns for early diagnostics. Dallas, TX: IEEE; 2013 Presented at: International Conference on Data Mining (ICDM); December 7-10, 2013; Dallas, TX URL: http://ieeexplore.ieee.org/document/6729504/
18. Wang F, Zhang P, Wang X, Hu J. Clinical risk prediction by exploring high-order feature correlations. AMIA Annu Symp Proc 2014;2014:1170-1179 [FREE Full text] [Medline: 25954428]
19. Wang F, Zhang P, Qian B, Wang X, Davidson I. Clinical risk prediction with multilinear sparse logistic regression. : ACM; 2014 Presented at: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14; August 24-27, 2014; New York, NY URL: http://dl.acm.org/citation.cfm?id=2623330 [doi: 10.1145/2623330.2623755]
20. Taslimitehrani V, Dong G, Pereira NL, Panahiazar M, Pathak J. Developing EHR-driven heart failure risk prediction models using CPXR(Log) with the probabilistic loss function. J Biomed Inform 2016 Apr;60:260-269 [FREE Full text] [doi: 10.1016/j.jbi.2016.01.009] [Medline: 26844760]
21. Lipton ZC, Kale DC, Elkan C, Wetzell R. Learning to Diagnose with LSTM Recurrent Neural Networks. Cornell University Library. 2015. URL: https://arxiv.org/abs/1511.03677 [accessed 2016-11-11] [WebCite Cache ID 6lvzKQRtY]
22. Tran T, Nguyen TD, Phung D, Venkatesh S. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). J Biomed Inform 2015;54:96-105 [FREE Full text]
23. Che Z, Kale D, Li W, Bahadori MT, Liu Y. Deep computational phenotyping. New York, NY: ACM; 2015 Presented at: KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 10-13, 2015; Sydney, NSW, Australia p. 507-516 URL: http://dl.acm.org/citation.cfm?id=2783365
24. Liu C, Wang F, Hu J, Xiong H. Temporal phenotyping from longitudinal electronic health records: a graph based framework. New York, NY: ACM; 2015 Presented at: KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 10-13, 2015; Sydney, NSW, Australia p. 705-714 URL: http://dl.acm.org/citation.cfm?id=2783352 [doi: 10.1145/2783258.2783352]
25. Rong X. word2vec Parameter Learning Explained. Cornell University Library. 2014. URL: https://arxiv.org/abs/1411.2738 [accessed 2016-11-09] [WebCite Cache ID 6ltrmTvYQ]
26. Breese JS, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. 1998 Presented at: UAI'98 Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence; July 24-26, 1998; Madison, WI p. 43-52 URL: http://dl.acm.org/citation.cfm?id=2074100
27. Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101:e215-e220 [FREE Full text] [doi: 10.1161/01.CIR.101.23.e215]
28. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biom J 2005;47:458-472 [FREE Full text] [doi: 10.1002/bimj.200410135]

Abbreviations

AUC: area under the receiver operating characteristic curve
CKD: chronic kidney disease
EHR: electronic health records
LSTM: long short-term memory
EMR: electronic medical records
fMRI: functional magnetic resonance imaging
LR: logistic regression
MIMIC-III: medical information mart for intensive care III
PDES: patient-diagnosis event similarity
PDPS: patient-diagnosis projection similarity
ROC: receiver operating characteristic
SVM: support vector machine

XSL•FO
RenderX

Edited by D Giordano; submitted 13.05.16; peer-reviewed by J Devenport, H Lehmann; comments to author 12.06.16; revised version received 05.08.16; accepted 02.11.16; published 25.11.16.

Please cite as:
Farhan W, Wang Z, Huang Y, Wang S, Wang F, Jiang X
A Predictive Model for Medical Events Based on Contextual Embedding of Temporal Sequences
JMIR Med Inform 2016;4(4):e39
URL: http://medinform.jmir.org/2016/4/e39/
doi:10.2196/medinform.5977
PMID:27888170

©Wael Farhan, Zhimu Wang, Yingxiang Huang, Shuang Wang, Fei Wang, Xiaoqian Jiang. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 25.11.2016. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

It's Time for Innovation in the Health Insurance Portability and Accountability Act (HIPAA)

Karen Colorafi[1*], PhD, RN; Bryan Bailey[2*], JD

[1]College of Nursing, Washington State University, Spokane, WA, United States

[2]Millligan Lawless, Phoenix, AZ, United States

[*]all authors contributed equally

Corresponding Author:
Karen Colorafi, PhD, RN
College of Nursing
Washington State University
103 E Spokane Falls Blvd
Spokane, WA, 99202
United States
Phone: 1 509 324 7318
Fax: 1 509 324 7341
Email: karen.colorafi@wsu.edu

Abstract

Whether it is the result of a tragic news story, a thoughtful commentary, or a segment on the entertainment networks, patient privacy rights are never far from the top of our minds. The Privacy and Security Rules contained in the Health Insurance Portability and Accountability Act of 1996 (HIPAA) represent a concerted effort to protect the privacy and security of the volumes of patient data generated by the health care system. However, the last twenty years has seen innovations and advancements in health information technology that were unimaginable at that time. It is time for innovation to the Privacy and Security Rules. We offer a common and relatable scenario as proof that certain Privacy and Security Rules can tie the hands of educators and innovators and need to be transformed.

(JMIR Med Inform 2016;4(4):e34) doi:10.2196/medinform.6372

KEYWORDS

innovation; HIPAA; electronic health record demonstration

Introduction

Recently we came across an art exhibit hosted by a prestigious American school in which a portable printer was set to download the messages sent through a hospital's digital pager system. We understand the artist stumbled upon the messages innocently one day while scanning various radio frequencies. The realization that pager data was so easily accessible prompted the artist to create the unique installation. This bold and creative act calls our attention to the abundance of intricate technology in our health care system, the lack of intention to the unintended consequences of its use, and the need we have to deploy technology safely. In other words, there is an innovation gap in play.

Weiss and Legrand (2011) define the innovation gap as the difference between the stated importance of innovation and the actual results achieved in an organization [1]. In its day, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) represented a significant advance: society's commitment to the protection of patient data, defending their rights by keeping sensitive health care information private and secure. Over a decade later, the Health Information Technology for Economic and Clinical Health (HITECH) Act [2] included in the American Recovery and Reinvestment Act (ARRA) stimulus legislation (2009) acknowledged some of the technological enhancements associated with the science of health care delivery and increased the penalties associated with violating the Act in a collective effort to promote the proper guardianship of health care data (Department of Health and Human Services, DHHS) [3]. The Security Rule was created with unusual foresight as a set of flexible requirements that could change and adapt with innovation.

Yet every week, the headlines online and in the papers discuss significant HIPAA infractions. The US Office for Civil Rights maintains a website dedicated to the public reporting of breaches affecting 500 or more individuals [4]. Online bloggers have

XSL•FO
RenderX

publicly questioned whether details leaked to the press about the circumstances surrounding the recent death of the artist Prince constituted a HIPAA violation [5], illustrating the heightened anxiety the general public feels about the ability of the health care profession to adequately protect the privacy and security of health care data. The scholarly literature continues to report that concerns about data breaches is a chief concern of patients, ultimately affecting the trust a patient places in a provider and in a health care facility [6]. We listen to stories from our friends and patients about the battles they have mounted to gain access to their own health care data.

We wrestle within our own organizations to make sense of HIPAA and to deploy its requirements responsibly while rolling out the next generation of health information technology (HIT), like real-time clinical dashboards and apps. Some have argued the iniquitousness of a rule that applies to health care apps but not consumer apps, even when they contain similar information [7-9]. We struggle to train a new generation of health care providers on electronic health record (EHR) systems and we refuse to share data with researchers out of fear of violating the rules. In short, it seems at times that our use of the Privacy and Security Rules has not adapted or supported the achievements and demands of health care.

We propose that health care leaders consider the significance of the innovation gap by deliberating a common scenario, one encountered by the authors on a regular basis: the EHR demonstration. Leaders in health care facilities who are justifiably proud of their EHR system are often approached by colleagues, educators, and vendor prospects to give demonstrations. Demonstrations are conducted for a variety of purposes: to show a colleague something that is especially fantastic or problematic with a particular system, to train health care providers, clinicians, or support staff, or to close a big sale. While the opportunity to showcase a beautiful system seems like a helpful thing to do – a professional courtesy of sorts – the facility ("covered entity") ought to carefully consider its responsibilities under the Act before agreeing to provide a demonstration.

Recently, one of the authors attended three different EHR demonstrations alongside a group of health care administration graduate students. Each of the demonstrations was given in a live production database and two out of the three used real patient encounters to demonstrate various scheduling, registration, billing, and clinical documentation scenarios. One student whose wife was a patient in one of the practices spent the entire session overwhelmed with anxiety that the next record revealed would be one with which he was intimately familiar.

This viewpoint provides health care leaders with a short review of HIPAA essentials, offers a compelling scenario suggesting the need for innovation, and provides suggested approaches to protecting patient privacy, working within the current confines of the HIPAA Privacy and Security Rules.

What is Protected Health Information?

Protected health information (PHI) includes all individually identifiable health information held or transmitted by a covered entity (or its business associates) in any form. Individually, identifiable health information is that which is created or received by a health care provider, health plan, employer, or health care clearinghouse which (1) relates to the past, present, or future physical or mental health or condition of an individual, the provision of health care to an individual, or the past, present, or future payment for the provision of health care to an individual; and (2) either identifies the individual or can be used to identify the individual. Electronic protected health information (e-PHI) is PHI that is maintained or transmitted in an electronic media, such as an EHR or practice management system and is afforded the same protections.

Why Do I Have to Protect It?

The Privacy Rule prohibits covered entities from using and disclosing PHI (including e-PHI), except as permitted or required by the Rule. The Security Rule requires covered entities to maintain reasonable and appropriate administrative, technical, and physical safeguards to protect e-PHI. For example, a covered entity must ensure the confidentiality of, anticipate threats to, and protect against impermissible uses and disclosures of e-PHI that resides in an EHR or practice management system by using safeguards such as complex and changing passwords, firewalls, and locking the server room. Failing to comply with the Privacy or Security Rule may result in civil monetary and criminal penalties. In addition, violations of the Privacy Rule may require written notifications of the impermissible use or disclosure to the affected individual(s), the Office for Civil Rights, and the media.

When Can Protected Health Information Be Used or Disclosed?

Generally speaking, the Privacy Rule prohibits covered entities from using or disclosing an individual's PHI without first obtaining the individual's prior written authorization. However, there are a number of exceptions to this Rule (Textbox 1).

Textbox 1. Exceptions to the Privacy Rule.

1. Giving information to the individual.

2. For treatment, payment, and health care operations (see [10] for a quick definition of these terms).

3. To persons involved in the individual's care after providing the individual with an opportunity to verbally agree or object, except in emergencies (eg, using the individual's name in a facility directory, paying a spouse's bill, and picking up a prescription for a family member).

4. Incidental disclosures of PHI resulting from a permitted use or disclosure (eg, a person glimpses another patient's name on a sign-in sheet).

5. For certain public interest purposes, such as disclosures that are required by law (eg, communicable diseases or child abuse).

What Can Be Disclosed?

At best, it is unclear whether a covered entity can disclose PHI during a demonstration. If a health care facility was under investigation for a violation, you might retrospectively argue that the Privacy Rule’s definition of “health care operations” includes “training programs in which students, trainees, or practitioners in areas of health care learn under supervision to practice or improve their skills as health care providers” and “training of non-health care professionals”. After all, a reasonable person may question how we plan to adequately train a new generation of programmers, information technology professionals, data scientists, business administration, and clinical students without acquainting them with one of health care’s most powerful tools. However, we would *not* prospectively advise a covered entity to disclose PHI during a demo based on this argument. Even if the disclosure is permitted, the covered entity still must comply with the Privacy Rule’s minimum necessary and reasonable safeguard requirements, which means the covered entity must have reasonable safeguards in place to ensure it only discloses the minimum PHI necessary for the demo. This is easier said than done. For these reasons, practices are safer by not disclosing any PHI during a demo. In addition, since there is a risk of improper and incidental disclosures of PHI while the demo participants are in your office, you must ensure that safeguards are in place to minimize these risks.

What Should I Do?

Tips that can help you prepare for a satisfying EHR demonstration while fulfilling your obligations under the Privacy and Security Rules are shown in Textbox 2.

It is important to remember that innovation does not simply happen once. A learning organization will revisit their policies and procedures related to the protection of data at least annually, or when a change in infrastructure demands (another requirement of the Act). Furthermore, we ought to consider that an Act that was innovative in 1996 may no longer solve the problems it was created to address, partly because the nature of the problem has changed. Academia has a desperate need to train students on the optimal use of EHR and practice management systems, which are commonplace across the country and represent the new standard of care. Health care businesses have an urgent need to partner with professionals and scholars who can analyze and make sense of their own EHR data. Industry could innovate and invent solutions to pressing and costly problems with adequate access to information. However, health professions training, big data, pharmacogenetics, and the re-selling of health care datasets are issues scantly addressed by the Act. We are well served to remember that innovation is best thought of as a process, not an outcome, that occurs within social environments that are dynamic and constantly changing [11]. We posit that health care needs innovation in the Privacy and Security rules to address the complexity that is inherent within the system in which we work and seek care.

Textbox 2. Tips to help you prepare for a satisfying EHR demonstration.

- Develop a policy and procedure for your HIPAA Privacy and Security set to explain the rules governing demonstrations of your EHR or practice management systems.
- Educate staff on the Privacy and Security Rules and your privacy and security policies and procedures (eg, be clear about what constitutes PHI, such as names on schedules).
- Always demo out of a test, build, or train (non-production) database.
- Ensure that the demo database does not contain actual PHI (sometimes configuration databases are back-loaded with real patient data from the live system).
- If you do not have a unique demo database:
  - Make sure there is a unique demo user login to your production database that does not have access to live patient data (eg, tasks, documents, and labs to review) and instead, demo test patients (eg, Donald Duck, James Cerner, and Abbey Allscripts).
  - Consider preparing a demo using screen shots (PHI redacted) on PowerPoint slides instead of using your production EHR. This is especially effective with “live” technologies such as telemedicine systems or state-run drug database inquiries.
- If appropriate to your situation, ensure your guests have signed a business associate agreement.
- Keep a log of dates and times when demos were provided and the names of attendees.
- Ask attendees to put mobile phones and tablets (eg, devices with cameras) in a basket before the demo begins and give them back when the demo is complete.

Acknowledgments

The authors wish to acknowledge their colleagues with whom animated debate occurred around the ethics of the EHR demonstration and need for change, including Cathy Lalley, Kathy Malloch, and Dan Simonson.

XSL•FO
RenderX

Conflicts of Interest

None declared

References

1. Weiss D, Legrand C. Innovative Intelligence: The Art and Practice of Leading Sustainable Innovation in Your Organization. Mississauga, Ontario: John Wiley & Sons Canada, Ltd; 2011.
2. U.S. Department of Health and Human Services. HITECH Act Enforcement Interim Final Rule. 2009 Nov 30. URL: http://www.hhs.gov/hipaa/for-professionals/special-topics/HITECH-act-enforcement-interim-final-rule/index.html [accessed 2016-10-20] [WebCite Cache ID 6lP6khm3C]
3. U.S. Department of Health and Human Services. Health Information Privacy. URL: http://www.hhs.gov/hipaa [accessed 2016-10-21] [WebCite Cache ID 6lQbnfujg]
4. U.S. Department of Health and Human Services. Breach portal: notice to the Secretary of HHS breach of unsecured protected health information. URL: https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf [accessed 2016-10-20] [WebCite Cache ID 6lPC5KlZ9]
5. Sivilli F. Leaks about Prince's death might be HIPAA violations. MedCity News. 2016 Apr 25. URL: http://medcitynews.com/2016/04/prince-hipaa/ [accessed 2016-10-20] [WebCite Cache ID 6lPCDnaMu]
6. Agaku IT, Adisa AO, Ayo-Yusuf OA, Connolly GN. Concern about security and privacy, and perceived control over collection and use of health information are related to withholding of health information from healthcare providers. J Am Med Inform Assoc 2014;21(2):374-378 [FREE Full text] [doi: 10.1136/amiajnl-2013-002079] [Medline: 23975624]
7. Downey R. Telemedicine and HIPAA compliancy. GlobalMed Telehealth Answers Blog. URL: https://www.globalmed.com/telehealthanswers/telemedicine-and-hipaa-compliancy [accessed 2016-10-20] [WebCite Cache ID 6lPCRZCO4]
8. Quintini H, Cox HA. Digital health care alert: is your health care app subject to HIPAA? Fenwick & West, LLP. 2016 Apr 05. URL: https://www.fenwick.com/publications/pages/is-your-health-care-app-subject-to-hipaa.aspx [accessed 2016-10-20] [WebCite Cache ID 6lPCapKY2]
9. Rosenfeld S. IRBs and big-data research—we’re aLL confused, Part 1. Quorum. 2016 Jul 11. URL: http://www.quorumreview.com/irbs-and-big-data-research-were-all-confused/?utm_source= mstr-list&utm_medium=email&utm_campaign=dr-rosenfeld-blog&utm_content= link-wired-big-data-part-1-07-12-16&utm_term=?link=headerlogo&mkt_tok= eyJpIjoiTmpZMFl6WTFPVEJqT0RrMyIsInQiOiJuUVJKM25EbW1pT1N4RHhIRWQyVUNlN1hcL0ZqQm1mMWExV0xJcFpkXC9NQW13WG9Vb0RCOEo3SklLT3gzWW1oRG5BQ01uSHA0MmdTOUI0SVh0dlh4cStKSUsyQUcyT1loTGpNK05CSVVVcmthTmFTTUFjc2RhclNSZUthYzlhZXQyIn0%3D [WebCite Cache ID 6lfkgzMIT]
10. HIPAA Survival Guide. HIPAA privacy rule 164.506. 2002 Aug 14. URL: http://www.hipaasurvivalguide.com/hipaa-regulations/164-506_BAK_01202013.php [accessed 2016-10-20] [WebCite Cache ID 6lPCwxlIl]
11. Fonseca J. Complexity and Innovation in Organizations. New York, NY: Routledge; 2002.

Abbreviations

EHR: electronic health record
e-PHI: electronic protected health information
HIPAA: Health Insurance Portability and Accountability Act
PHI: protected health information

Edited by G Eysenbach; submitted 19.07.16; peer-reviewed by I Brooks, K Malloch; comments to author 17.08.16; revised version received 20.08.16; accepted 21.09.16; published 02.11.16.

Please cite as:
Colorafi K, Bailey B
It’s Time for Innovation in the Health Insurance Portability and Accountability Act (HIPAA)
JMIR Med Inform 2016;4(4):e34
URL: http://medinform.jmir.org/2016/4/e34/
doi:10.2196/medinform.6372
PMID:27806923

©Karen Colorafi, Bryan Bailey. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 02.11.2016. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Natural Language Processing–Enabled and Conventional Data Capture Methods for Input to Electronic Health Records: A Comparative Usability Study

David R Kaufman[1], PhD; Barbara Sheehan[2], NP, PhD; Peter Stetson[3], MA, MD; Ashish R Bhatt[4], MBA; Adele I Field[4], MFA; Chirag Patel[5], MD, PhD; James Mark Maisel[4], MD

[1]Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ, United States

[2]Health Strategy and Solutions, Intel Corp, Santa Clara, CA, United States

[3]Internal Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, United States

[4]ZyDoc Medical Transcription LLC, Islandia, NY, United States

[5]Department of Neurology & Neurological Sciences, Stanford School of Medicine, Stanford University, Palo Alto, CA, United States

Corresponding Author:
David R Kaufman, PhD
Department of Biomedical Informatics
Arizona State University
13212 East Shea
Scottsdale, AZ, 85260
United States
Phone: 1 4808840250
Fax: 1 480 884 0239
Email: dave.kaufman@asu.edu

Abstract

Background: The process of documentation in electronic health records (EHRs) is known to be time consuming, inefficient, and cumbersome. The use of dictation coupled with manual transcription has become an increasingly common practice. In recent years, natural language processing (NLP)–enabled data capture has become a viable alternative for data entry. It enables the clinician to maintain control of the process and potentially reduce the documentation burden. The question remains how this NLP-enabled workflow will impact EHR usability and whether it can meet the structured data and other EHR requirements while enhancing the user's experience.

Objective: The objective of this study is evaluate the comparative effectiveness of an NLP-enabled data capture method using dictation and data extraction from transcribed documents (NLP Entry) in terms of documentation time, documentation quality, and usability versus standard EHR keyboard-and-mouse data entry.

Methods: This formative study investigated the results of using 4 combinations of NLP Entry and Standard Entry methods ("protocols") of EHR data capture. We compared a novel dictation-based protocol using MediSapien NLP (NLP-NLP) for structured data capture against a standard structured data capture protocol (Standard-Standard) as well as 2 novel hybrid protocols (NLP-Standard and Standard-NLP). The 31 participants included neurologists, cardiologists, and nephrologists. Participants generated 4 consultation or admission notes using 4 documentation protocols. We recorded the time on task, documentation quality (using the Physician Documentation Quality Instrument, PDQI-9), and usability of the documentation processes.

Results: A total of 118 notes were documented across the 3 subject areas. The NLP-NLP protocol required a median of 5.2 minutes per cardiology note, 7.3 minutes per nephrology note, and 8.5 minutes per neurology note compared with 16.9, 20.7, and 21.2 minutes, respectively, using the Standard-Standard protocol and 13.8, 21.3, and 18.7 minutes using the Standard-NLP protocol (1 of 2 hybrid methods). Using 8 out of 9 characteristics measured by the PDQI-9 instrument, the NLP-NLP protocol received a median quality score sum of 24.5; the Standard-Standard protocol received a median sum of 29; and the Standard-NLP protocol received a median sum of 29.5. The mean total score of the usability measure was 36.7 when the participants used the NLP-NLP protocol compared with 30.3 when they used the Standard-Standard protocol.

Conclusions: In this study, the feasibility of an approach to EHR data capture involving the application of NLP to transcribed dictation was demonstrated. This novel dictation-based approach has the potential to reduce the time required for documentation

XSL•FO
RenderX

and improve usability while maintaining documentation quality. Future research will evaluate the NLP-based EHR data capture approach in a clinical setting. It is reasonable to assert that EHRs will increasingly use NLP-enabled data entry tools such as MediSapien NLP because they hold promise for enhancing the documentation process and end-user experience.

(JMIR Med Inform 2016;4(4):e35) doi:10.2196/medinform.5544

KEYWORDS

electronic health records; natural language processing; medical transcription; user-computer interface

Introduction

Electronic Health Records and Data Entry

Electronic health records (EHRs) permeate most medical practices in the United States [1]. A promising feature of EHRs is that they provide machine-readable structured data that can be stored electronically, so that patient-centered information can be reviewed, retrieved, reported, and shared in real time to facilitate patient care. Although narrative data entry supports a measure of flexibility, structured data entry confers a number of advantages such as ready access to clinical decision support and interoperability between EHRs and health information exchanges. To achieve the full complement of these benefits, health care providers must generate clinical notes and reports in both human-readable and machine-readable formats. This adds effort to the documentation workflow [2,3] and requires new computer skills of physicians. The additional work required has led to a growing number of data entry alternatives [4], which is the subject of this paper.

Since their inception, EHRs “have been proposed as a means for improving availability, legibility, and completeness of patient information” [5]. The potential benefits of EHRs as instruments of patient care are widely recognized. Spurred by the 2009 US Health Information Technology for Economic and Clinical Health (HITECH) Act and accompanying incentives for providers to use EHRs, advancements in EHR technologies and implementation in the United States have grown rapidly. Approximately 78% of office-based physicians reported using some form of EHR in 2013 [6]. The role of EHRs is now considered integral to achieving federal health care goals, as expressed in the Meaningful Use mandate, for example. This has compelled physicians to adapt to new methods of documentation with concomitant changes to clinical workflow. This has resulted in great uncertainty about the impact of these requirements on the effective application of EHR systems [7].

As EHR implementations continue, physicians frequently express dissatisfaction with EHR documentation methods and usability [8]. Problems associated with usability impact not only the quality of patient records but can even contribute to compromised patient safety [9,10]. EHR documentation places ever-increasing demands on clinicians’ time, which contributes further to diminished quality of documents (eg, replete with irrelevant, redundant, and erroneous information) and physician dissatisfaction. EHR usability is a complex problem involving a multitude of factors [11-13], many of which are beyond the scope of this study. The focus in this paper is on the usability of data capture methods designed to enhance and potentially alleviate some of the burden resulting from manual input methods.

Natural Language Processing–Based Solutions

Natural language processing (NLP) has emerged as a viable solution for clinical data capture. Many challenges remain for keyboard-and-mouse entry, namely, having to type text and negotiate the often unwieldy EHR interface to record information in structured fields. This is exacerbated by the fact that much of the EHR content continues to be unstructured [3,14]. A 2015 American Medical Informatics Association report identified time-consuming data entry as a problem with EHRs and recommended improving the EHR interface by allowing “multiple modes of data entry to accommodate provider preferences, including voice, typing, clicking, and handwriting recognition” [15].

Although most clinical information in EHRs is stored as unstructured data, such as clinical narrative, its electronic capture or retrieval has been challenging [16]. NLP has the potential to enable the clinician to reduce the documentation burden with the advantages of dictation—efficiency, usability, and quality—and also satisfy the needs for both machine-readable structured data and human-usable rich text in the EHR.

Problems associated with the time required for documentation and usability are well established. However, there is also evidence to suggest that quality of EHR documents (eg, progress reports) is problematic [17,18]. Physician documents often contain redundant, extraneous information or missing and inaccurate patient data [17]. EHR notes are not optimally used to either facilitate clinical communication or enhance patient care [8]. The measure of the quality and completeness of data in the EHR represents a challenging issue [19,20]. Stetson and colleagues [21] developed a tool for quantifying documentation quality, the Physician Documentation Quality Instrument (PDQI), and demonstrated its construct validity and internal consistency reliability. The initial tool consisted of 22 items and was subsequently reduced to a 9-item tool in order to facilitate its real-world application. The instrument can be used to assess the output quality of EHR note modules, as well as explain the components of document areas in need of improvement. Stetson and colleagues [21] assessed the interrater reliability using the intraclass correlation for consistency of average measures on the PDQI-9 total scores and found it to be 0.83 (CI 0.72-0.91). The tool can reliably be used to compare documentation methods and changes in quality resulting from a change in such methods.

In this study, we were focally concerned with testing NLP-enabled dictation-based data capture as a potential solution for relieving the increased burden of documentation. The benchmarks of performance include measures of time, data

XSL•FO RenderX

quality, and usability. According to Cimino [5], “Improvements in the documentation process hold promise for more than simply reduced data entry effort and more readable notes. If impressions and plans can be captured as explicit data elements, using standard terminology, rather than being buried in the narrative text of a note, EHRs could use this information to better support clinical work flow.” As a result of physicians capturing explicit data elements, their clinical reasoning can be made more transparent and more easily available to colleagues caring for the same patient via electronic access to their EHR or data exchange.

Data Capture Methods

A variety of modalities have been used for creating clinical documentation for EHR data capture or extraction to generate structured, actionable data (ie, data that are consumable, usable, reusable by a computer, and exchangeable with other computer systems in an efficient manner). These modalities include paper-based records transfer; verbal communication; direct entry or direct entry with macros; electronic templates; “Smart Forms”; dictation using speech recognition, sometimes known as voice recognition or continuous speech recognition; transcription or transcription with manual error correction; patient-recorded data (various methods); and hybrids, with or without NLP data capture, also termed “text processing” [22]. Rosenbloom et al observe that in spite of a “profusion of computer-based documentation (CBD) systems that promote real-time structured documentation,” it is a challenge “integrating clinical documentation into workflows that contain EHR systems.” They further note that health care providers prefer the ability to achieve a certain balance by both using a standardized note structure and having the flexibility to use expressive narrative text, facilitated by speech recognition. NLP systems afford that expressivity in developing a patient narrative as well as offering the capability to encode structured notes in a range of clinical document types and forms [22,23].

Figure 1 illustrates 5 alternative dictation-based EHR data capture methods. The NLP Entry method used in the study is shown in the center of the figure (labeled as 3), with bold arrows and boxes. Methods 1 and 4 show speech recognition and transcription, respectively, converting dictation to text that is inserted into the EHR. Methods 2 and 3 show NLP being applied to the speech-recognized and transcribed text, respectively, to generate structured data that are inserted into the EHR alongside the text. Method 5 shows a human scribe manually entering text and structured data into the EHR immediately in live time as the physician dictates or at a later time from recorded dictation.

NLP encompasses a family of methods for processing text. These methods have been used for a range of EHR applications [24] including information extraction [25], information retrieval [26], question answering, and text summarization [27]. NLP has also been used for the automatic encoding of narrative text into EHRs [4,28]. NLP and associated technologies used in conjunction with dictation for capturing and structuring medical data have advanced considerably in recent years [4,29].

Whereas relatively few NLP systems for structured clinical data capture are implemented outside academic medical centers [22], NLP is gaining more traction as a viable commercial technology for populating EHRs. In addition to the MediSapien NLP (ZyDoc Medical Transcription LLC) application used in this study (the user interface for which is shown in Figure 2), there are a limited number of other NLP products being marketed or developed for use with EHR systems to enable NLP Entry (ie, free-text data capture, structuring, and EHR population). For example, both M*Modal [30] and Nuance [31] offer dictation products with voice recognition that structure some data for EHRs. Certain EHR vendors, such as Allscripts, Greenway, and Cerner among others, have integrated the M*Modal or Nuance technologies into their EHRs [32]. Other NLP-based research studies (including one on the interpretation of free-text Papanicolaou test reports for clinical decision support [33] and another on the use of “cognitive analytic tools to gain insight from all types of healthcare information,” including “knowledge-driven decision support” and “data-driven decision support” [34]) demonstrate the increasing importance of NLP in generating and analyzing structured health data.

The NLP engine used by the MediSapien NLP data capture application is the Medical Language Extraction and Encoding System (MedLEE), which was developed at Columbia University in the Department of Biomedical Informatics. MedLEE accepts unstructured clinical text inputs and outputs structured clinical information in a variety of formats [35]. Utilizing clinical lexicons, it is able to normalize clinical concepts in the text to conform to various standard terminologies. It is also able to identify, among other attributes, negation, degrees of certainty, temporal data, and results associated with the identified clinical concepts. MedLEE has been used for a number of data extraction purposes but was not specifically optimized for generating clinical documentation [36]. A commercially available version of MedLEE is now licensed and maintained by Health Fidelity under the product name REVEAL.

Developed by ZyDoc Medical Transcription, the MediSapien NLP data capture application allows doctors to use unstructured dictation to capture structured data in the EHR. MediSapien NLP preprocesses documents, leverages the MedLEE NLP engine, and postprocesses the NLP output using patent-pending processes that augment the NLP engine’s output. It also enables a workflow by which (1) the physician dictates, (2) the dictation is transcribed or subjected to speech recognition, (3) MediSapien NLP generates structured data from the transcription, and (4) the structured data and text are inserted into the EHR, although we simulated the EHR interface in the study.

Figure 2 shows part of a screen from the MediSapien NLP application in which source text is displayed on the left, with medical concepts highlighted, and structured data generated from the source text are displayed on the right. The document included in Figure 2 was selected to illustrate the volume of structured data generated by the MediSapien NLP application; the text was not produced as part of the study presented in this paper. As an indication of the volume of data generated by MediSapien NLP, the average number of clinical concepts and corresponding modifiers identified in a sample of notes from the study using the NLP-NLP protocol was 392. These concepts are coded in various standard terminologies—including ICD-10-CM (International Classification of Diseases, Tenth

XSL•FO
RenderX

Revision, Clinical Modification); ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification); SNOMED-CT (Systematized Nomenclature of Medicine, Clinical Terms); RxNorm; LOINC (Logical Observation Identifiers Names and Codes); and CPT (Current Procedural Terminology)—depending on the type of concept identified. Modifiers are structured data elements that provide additional properties related to a clinical concept. Examples of modifiers include body location, status, and dose (as shown on the right side of Figure 2).

It should be noted that this was a formative study designed to investigate the comparative effects of data capture methods enabled by the NLP system. The focus of the analysis was on characterizing interactive behavior and system usability rather than the NLP method. Future studies will investigate the efficacy of the NLP processes used by the system.

The objectives of this study were to (1) measure the effects, relative to using Standard Entry only, of using 3 NLP-based documentation protocols on EHR documentation time and quality and (2) measure the effects of an NLP Entry–based protocol and a Standard Entry–based protocol on the usability of the documentation process.

XSL•FO
RenderX

Figure 1. Five dictation-based electronic health record (EHR) data capture methods.

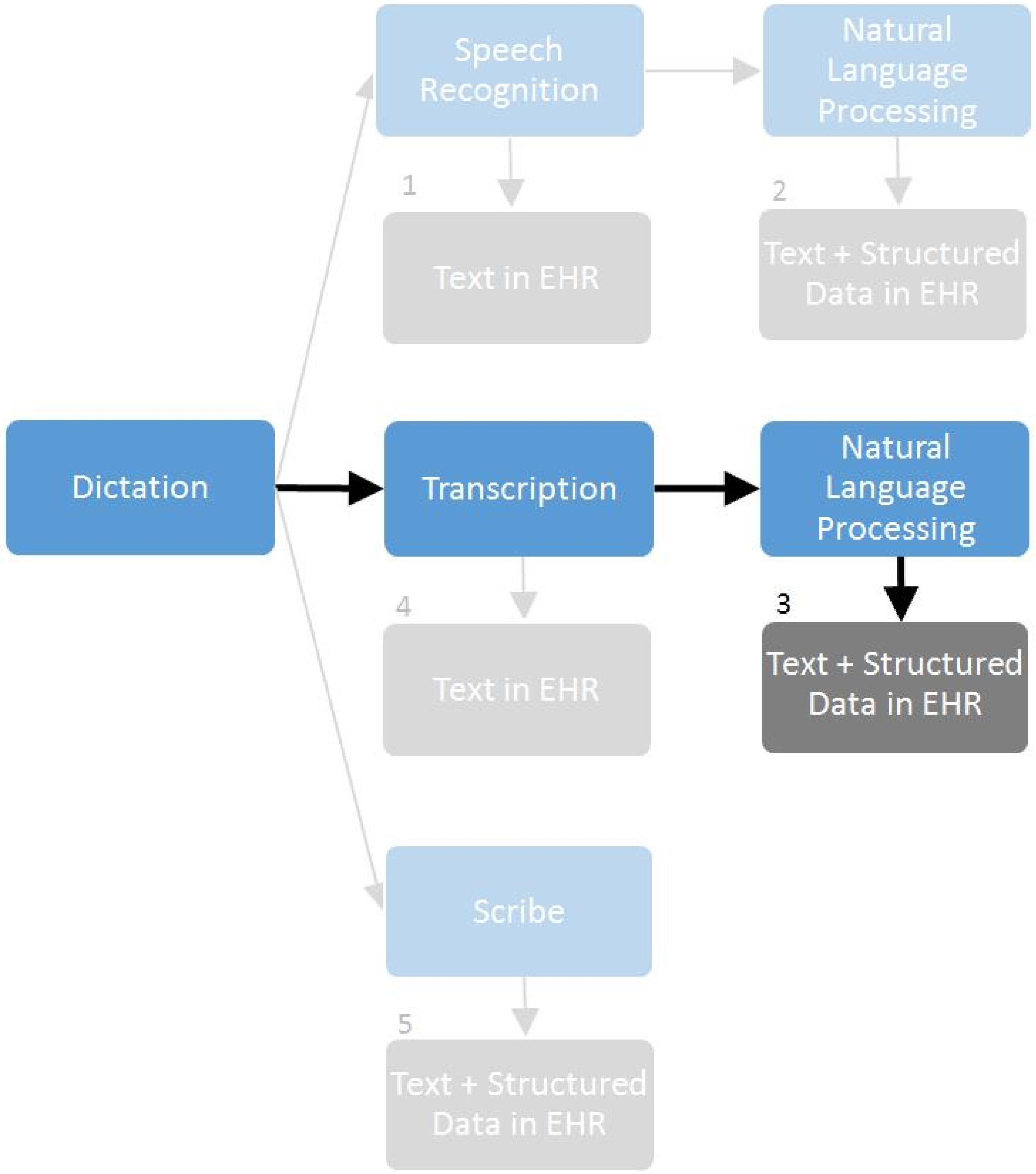
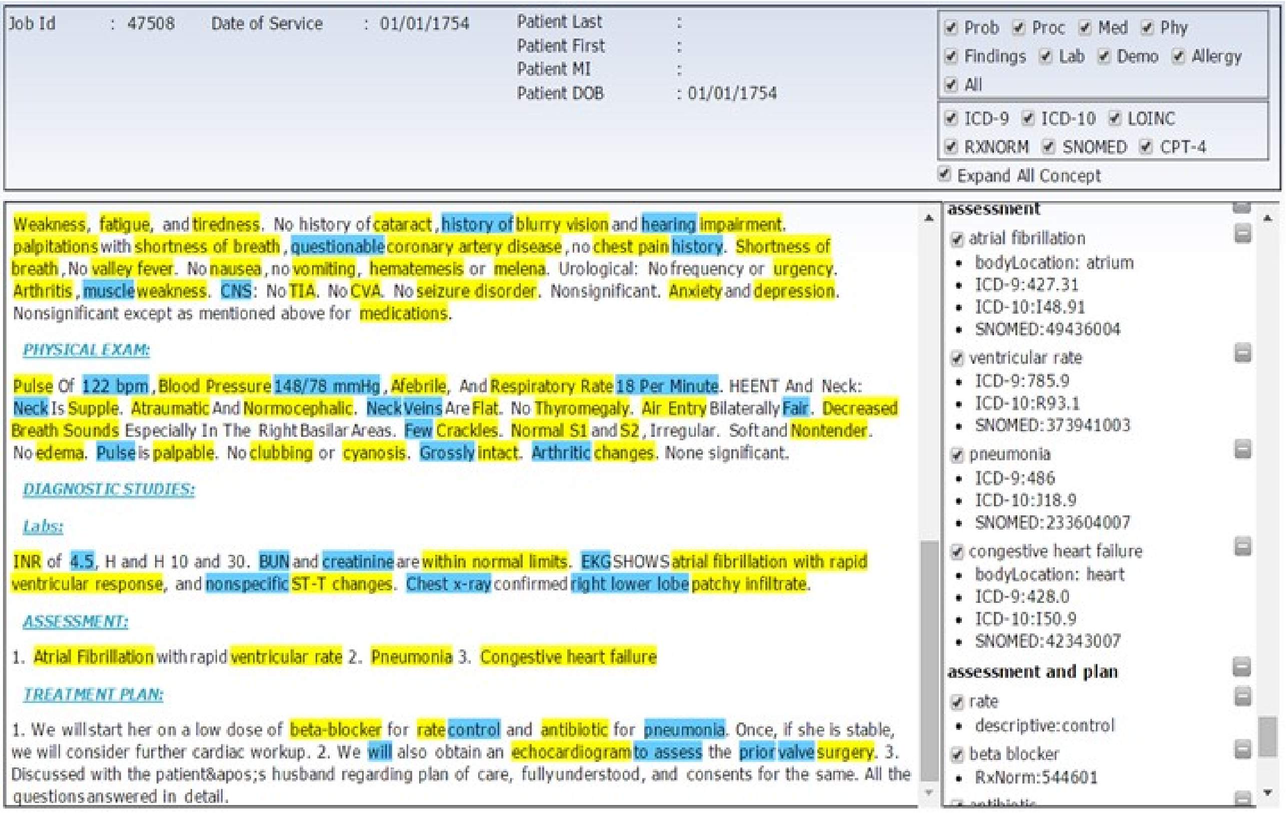

Speech Recognition
Natural Language Processing
1
Text in EHR
2
Text + Structured Data in EHR
Dictation
Transcription
Natural Language Processing
4
Text in EHR
3
Text + Structured Data in EHR
Scribe
5
Text + Structured Data in EHR

Figure 2. MediSapien NLP application user interface, illustrating the volume of structured data generated by MediSapien NLP. NLP: natural language processing.



Job Id : 47508 Date of Service : 01/01/1754

Patient Last :
Patient First :
Patient MI :
Patient DOB : 01/01/1754

☑ Prob ☑ Proc ☑ Med ☑ Phy
☑ Findings ☑ Lab ☑ Demo ☑ Allergy
☑ All

☑ ICD-9 ☑ ICD-10 ☑ LOINC
☑ RXNORM ☑ SNOMED ☑ CPT-4

☑ Expand All Concept

Weakness, fatigue, and tiredness. No history of cataract, history of blurry vision and hearing impairment. palpitations with shortness of breath, questionable coronary artery disease, no chest pain history. Shortness of breath, No valley fever. No nausea, no vomiting, hematemesis or melena. Urological: No frequency or urgency. Arthritis, muscle weakness. CNS: No TIA. No CVA. No seizure disorder. Nonsignificant. Anxiety and depression. Nonsignificant except as mentioned above for medications.

PHYSICAL EXAM:

Pulse Of 122 bpm, Blood Pressure 148/78 mmHg, Afebrile, And Respiratory Rate 18 Per Minute. HEENT And Neck: Neck Is Supple. Atraumatic And Normocephalic. Neck Veins Are Flat. No Thyromegaly. Air Entry Bilaterally Fair. Decreased Breath Sounds Especially In The Right Basilar Areas. Few Crackles. Normal S1 and S2, Irregular. Soft and Nontender. No edema. Pulse is palpable. No clubbing or cyanosis. Grossly intact. Arthritic changes. None significant.

DIAGNOSTIC STUDIES:

Labs:

INR of 4.5, H and H 10 and 30. BUN and creatinine are within normal limits. EKG SHOWS atrial fibrillation with rapid ventricular response, and nonspecific ST-T changes. Chest x-ray confirmed right lower lobe patchy infiltrate.

ASSESSMENT:

1. Atrial Fibrillation with rapid ventricular rate 2. Pneumonia 3. Congestive heart failure

TREATMENT PLAN:

1. We will start her on a low dose of beta-blocker for rate control and antibiotic for pneumonia. Once, if she is stable, we will consider further cardiac workup. 2. We will also obtain an echocardiogram to assess the prior valve surgery. 3. Discussed with the patient&apos;s husband regarding plan of care, fullyunderstood, and consents for the same. All the questions answered in detail.

assessment

☑ atrial fibrillation
- bodyLocation: atrium
- ICD-9:427.31
- ICD-10:I48.91
- SNOMED:49436004

☑ ventricular rate
- ICD-9:785.9
- ICD-10:R93.1
- SNOMED:373941003

☑ pneumonia
- ICD-9:486
- ICD-10:J18.9
- SNOMED:233604007

☑ congestive heart failure
- bodyLocation: heart
- ICD-9:428.0
- ICD-10:I50.9
- SNOMED:42343007

assessment and plan

☑ rate
- descriptive:control

☑ beta blocker
- RxNorm:544601

☑ antibiotic

Methods

Study Overview

The study evaluated an NLP-enabled solution for documentation. Specifically, we focused on three problem areas related to EHR data capture: (1) efficiency, including time required for data capture; (2) effectiveness, encompassing documentation quality; and (3) physician satisfaction, based on usability. We compared a novel dictation-based protocol using MediSapien NLP for structured data capture ("NLP-NLP") against a standard, keyboard-and-mouse structured data capture protocol where the study participant was instructed to generate EHR documentation as in normal clinical practice ("Standard-Standard") as well as 2 novel hybrid protocols ("NLP-Standard" and "Standard-NLP") to determine which protocols provided better results in terms of data capture time, documentation quality, and physician satisfaction. The hybrid protocols were included because we anticipated that mixed forms and modalities of interaction may serve as realistic alternatives to a one-dimensional NLP approach or standard data entry. For example, certain parts of clinical notes may be better served by one modality of entry or the other; a note's assessment and plan sections are often more given to free text and may therefore be suited to a dictation-based modality, whereas a note's history and physical examination sections are less so and therefore may be better suited to a different modality. A hybrid approach may offer greater flexibility and can be adapted to the preferences of individual users. The study presented here is formative work that focused more directly on the nature of user interaction and the user experience rather than the efficacy or precision of the NLP system or the system for insertion of data in the EHR. These will be addressed in the next phase of research.

Study Design

This study contrasted 4 conditions involving combinations of NLP Entry and Standard Entry (referred to in this paper as documentation protocols) on the following measures: documentation time, documentation quality, and usability of the documentation process.

The Standard Entry method (ie, how physicians typically use an EHR to document) entailed using a keyboard and mouse for typing text and negotiating the graphical user interface (eg, drop-down menus, check-boxes) to record information in structured fields.

In the NLP Entry method, the participants dictated the content of the documents. They did not enter any documentation using the keyboard or mouse. Their dictation was transcribed, and the transcription was inputted into the MediSapien NLP application. That application outputted a document containing structured data (an example of which is shown in Figure 2) generated from the transcription. Finally, following precise instructions, study assistants entered the transcribed text and part of the generated structured data into a Microsoft Word document to produce a final note.

XSL•FO
RenderX

Table 1. Documentation methods used for each documentation protocol.

Documentation protocols	Documentation method for history and physical examination	Documentation method for assessment and plan
NLP[a]-NLP	NLP Entry	NLP Entry
NLP-Standard	NLP Entry	Standard Entry
Standard-NLP	Standard Entry	NLP Entry
Standard-Standard (control)	Standard Entry	Standard Entry

[a] NLP: natural language processing.

In the study, each physician was asked to document 4 notes using 4 methods including 1 control method (Standard-Standard protocol) and 3 experimental protocols consisting of combinations of NLP Entry and Standard Entry for documenting different parts of the note, as presented in Table 1. The Standard-NLP protocol involved using Standard Entry to generate the history and physical examination sections and NLP Entry to generate the assessment and plan sections. The NLP-Standard protocol involved using NLP Entry for the history and physical examination sections and Standard Entry for the assessment and plan sections. The NLP-NLP protocol involved using NLP Entry for the entire note. The order in which the protocols were used was randomized for each participant.

Participants

Physician participants were recruited through referrals. Two of the coauthors (BS and PS) referred us to several physicians who in turn made additional referrals. The inclusion criteria for the participants were as follows: (1) must be a neurologist, cardiologist, or nephrologist, the 3 specialties included in the study; (2) must be a senior resident, fellow, or attending; and (3) must be a current user of the Columbia University Medical Center's (CUMC) Crown Allscripts EHR (Chicago, IL). The participants were each compensated US $500 for their efforts.

Setting

This study was conducted at CUMC. The test protocol was administered with physician participants at their offices. Fictitious patients were created for the study, and the participants documented their cases in a test environment of the Crown Allscripts EHR, which was the same EHR in which the participants documented during normal clinical practice. Participants were all experienced users of the system. The Crown Allscripts EHR had been in use in excess of 5 years at CUMC as of the time of the study. This test environment contained the same custom templates that participants used during normal clinical practice. As a result, the Standard Entry method simulated documentation during normal clinical practice as closely as reasonably possible.

Test Scripts

The test scripts were based on anonymized transcription documents that were modified by 4 expert clinicians (2 fellows and 2 attending physicians). These clinicians were not participants in the study. The test scripts consisted of history and physical examination sections but excluded assessment and plan sections. After reviewing test scripts that described cases of the fictitious patients, the participants generated 4 multisectional consultation or admission notes using 4 documentation protocols (Table 1).

Procedure

First, each participant read the instructions for generating consultation or admission notes based on the 4 provided test scripts, an example of which is shown in Figure 3. The instructions indicated that the participant must generate documentation without copying verbatim any part of the test script and that the assessment and plan sections of these notes would be generated based on the participant's medical judgment.

Second, the participant was asked to review the test scripts and to generate 4 notes from 4 test scripts, 4 examples of which are shown in Figures 4-7.

Third, for documentation in which NLP Entry was used, the participant's dictation was transcribed; the transcription was processed by MediSapien NLP; and the transcription, structured data generated, and any documentation generated for the note by Standard Entry (if applicable) were combined to create the final note. A simulated interface and simulated note were used for NLP Entry: following a protocol, study assistants copied the generated unstructured and structured data into a Microsoft Word document to generate the final note. For Standard Entry, an actual EHR interface was used.

Finally, after reviewing their final notes, the participants completed 2 System Usability Scale (SUS) surveys [37] to evaluate the usability of the NLP-NLP protocol and the Standard-Standard protocol. Given the limited availability of clinicians' time, we determined this would be the most important contrast to include in the study. The SUS is a widely used and reliable tool. It consists of 10 Likert items measured on a 5-point scale (ranging from "completely agree" to "completely disagree") [38]. Half of the items are framed as positive questions (eg, "easy to use") and half are negative (eg, "unnecessarily complex"). The scores were tabulated accordingly. The surveys were made available in SurveyMonkey, a Web-based survey application. The SUS was slightly modified for language and context. The questions with tabulated responses are presented in Table 2.

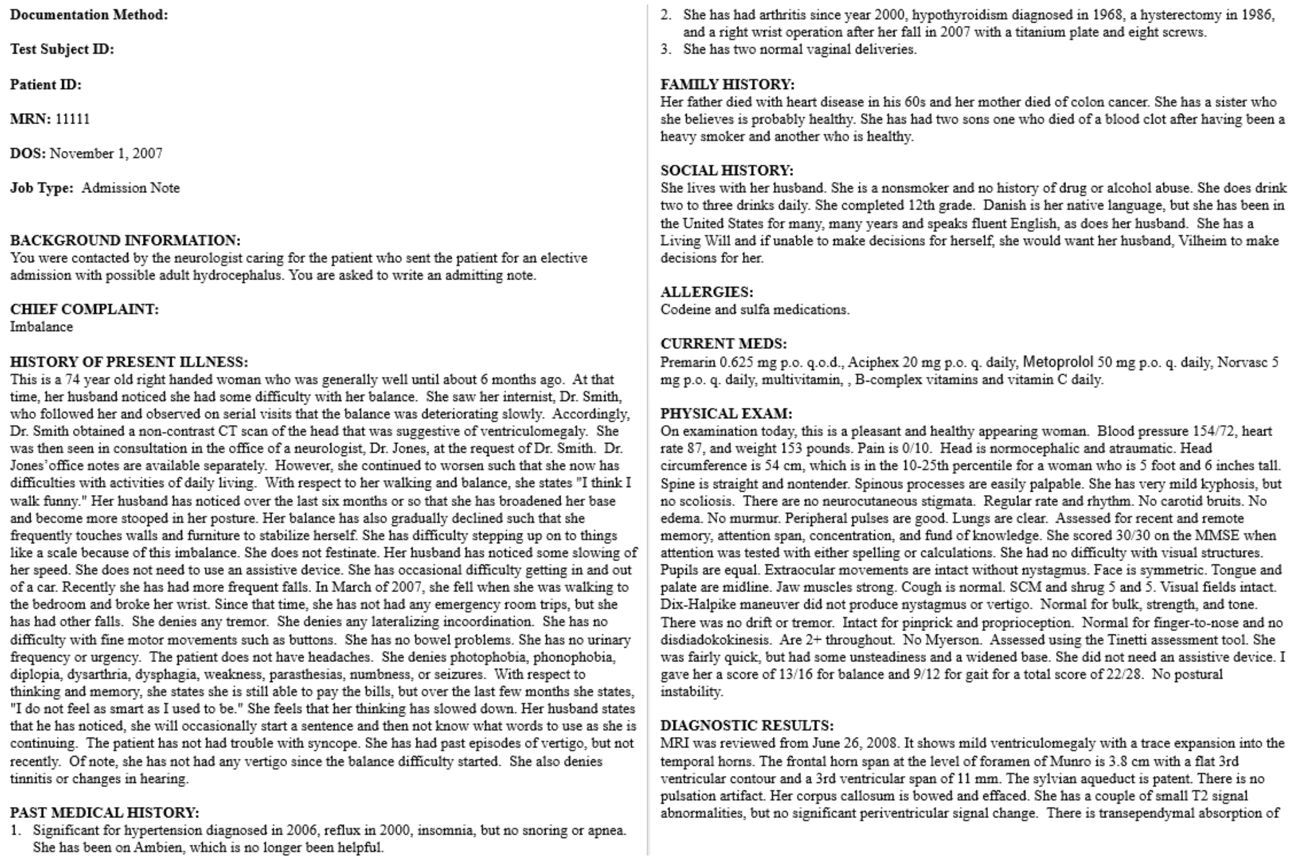
XSL•FO
RenderX

Table 2. Summary of usability scores (mean, SD) and paired *t* test comparisons between use of the Standard-Standard and the NLP-NLP protocols (n=23 cases); scores have been normalized such that higher scores indicate greater usability.

Usability question	Standard-Standard, mean (SD)	NLP[a]-NLP, mean (SD)	*P* value
I think that I would like to use this method frequently for admitting notes.	2.9 (0.9)	3.3 (0.8)	.21
I found this method unnecessarily complex.	2.5 (1.4)	3.8 (0.8)	.003
I thought this method was easy to use.	2.8 (1)	4.2 (0.6)	<.001
I think that I would need assistance to be able to use this method.	3.3 (1.1)	3.6 (0.9)	.24
I found the various functions in the processes of the method were well integrated.	2.6 (0.9)	3.2 (1)	.05
I would imagine that most people would learn to use this method very quickly.	3.0 (0.9)	3.8 (0.7)	.01
I found this method very cumbersome/awkward to use.	2.6 (1.1)	3.7 (0.9)	.004
I felt very confident using this method.	3.6 (0.8)	3.4 (0.8)	.43
I would need to learn a lot of things before I could get going with this method.	3.6 (1)	3.8 (0.8)	.40
I feel the method would fit well in my existing workflow.	2.8 (0.9)	3.4 (0.9)	.08

[a] NLP: natural language processing.

Figure 3. Example of a neurology test script.



Documentation Method:

Test Subject ID:

Patient ID:

MRN: 11111

DOS: November 1, 2007

Job Type: Admission Note

BACKGROUND INFORMATION:
You were contacted by the neurologist caring for the patient who sent the patient for an elective admission with possible adult hydrocephalus. You are asked to write an admitting note.

CHIEF COMPLAINT:
Imbalance

HISTORY OF PRESENT ILLNESS:
This is a 74 year old right handed woman who was generally well until about 6 months ago. At that time, her husband noticed she had some difficulty with her balance. She saw her internist, Dr. Smith, who followed her and observed on serial visits that the balance was deteriorating slowly. Accordingly, Dr. Smith obtained a non-contrast CT scan of the head that was suggestive of ventriculomegaly. She was then seen in consultation in the office of a neurologist, Dr. Jones, at the request of Dr. Smith. Dr. Jones'office notes are available separately. However, she continued to worsen such that she now has difficulties with activities of daily living. With respect to her walking and balance, she states "I think I walk funny." Her husband has noticed over the last six months or so that she has broadened her base and become more stooped in her posture. Her balance has also gradually declined such that she frequently touches walls and furniture to stabilize herself. She has difficulty stepping up on to things like a scale because of this imbalance. She does not festinate. Her husband has noticed some slowing of her speed. She does not need to use an assistive device. She has occasional difficulty getting in and out of a car. Recently she has had more frequent falls. In March of 2007, she fell when she was walking to the bedroom and broke her wrist. Since that time, she has not had any emergency room trips, but she has had other falls. She denies any tremor. She denies any lateralizing incoordination. She has no difficulty with fine motor movements such as buttons. She has no bowel problems. She has no urinary frequency or urgency. The patient does not have headaches. She denies photophobia, phonophobia, diplopia, dysarthria, dysphagia, weakness, parasthesias, numbness, or seizures. With respect to thinking and memory, she states she is still able to pay the bills, but over the last few months she states, "I do not feel as smart as I used to be." She feels that her thinking has slowed down. Her husband states that he has noticed, she will occasionally start a sentence and then not know what words to use as she is continuing. The patient has not had trouble with syncope. She has had past episodes of vertigo, but not recently. Of note, she has not had any vertigo since the balance difficulty started. She also denies tinnitis or changes in hearing.

PAST MEDICAL HISTORY:

1. Significant for hypertension diagnosed in 2006, reflux in 2000, insomnia, but no snoring or apnea. She has been on Ambien, which is no longer been helpful.
2. She has had arthritis since year 2000, hypothyroidism diagnosed in 1968, a hysterectomy in 1986, and a right wrist operation after her fall in 2007 with a titanium plate and eight screws.
3. She has two normal vaginal deliveries.

FAMILY HISTORY:
Her father died with heart disease in his 60s and her mother died of colon cancer. She has a sister who she believes is probably healthy. She has had two sons one who died of a blood clot after having been a heavy smoker and another who is healthy.

SOCIAL HISTORY:
She lives with her husband. She is a nonsmoker and no history of drug or alcohol abuse. She does drink two to three drinks daily. She completed 12th grade. Danish is her native language, but she has been in the United States for many, many years and speaks fluent English, as does her husband. She has a Living Will and if unable to make decisions for herself, she would want her husband, Vilheim to make decisions for her.

ALLERGIES:
Codeine and sulfa medications.

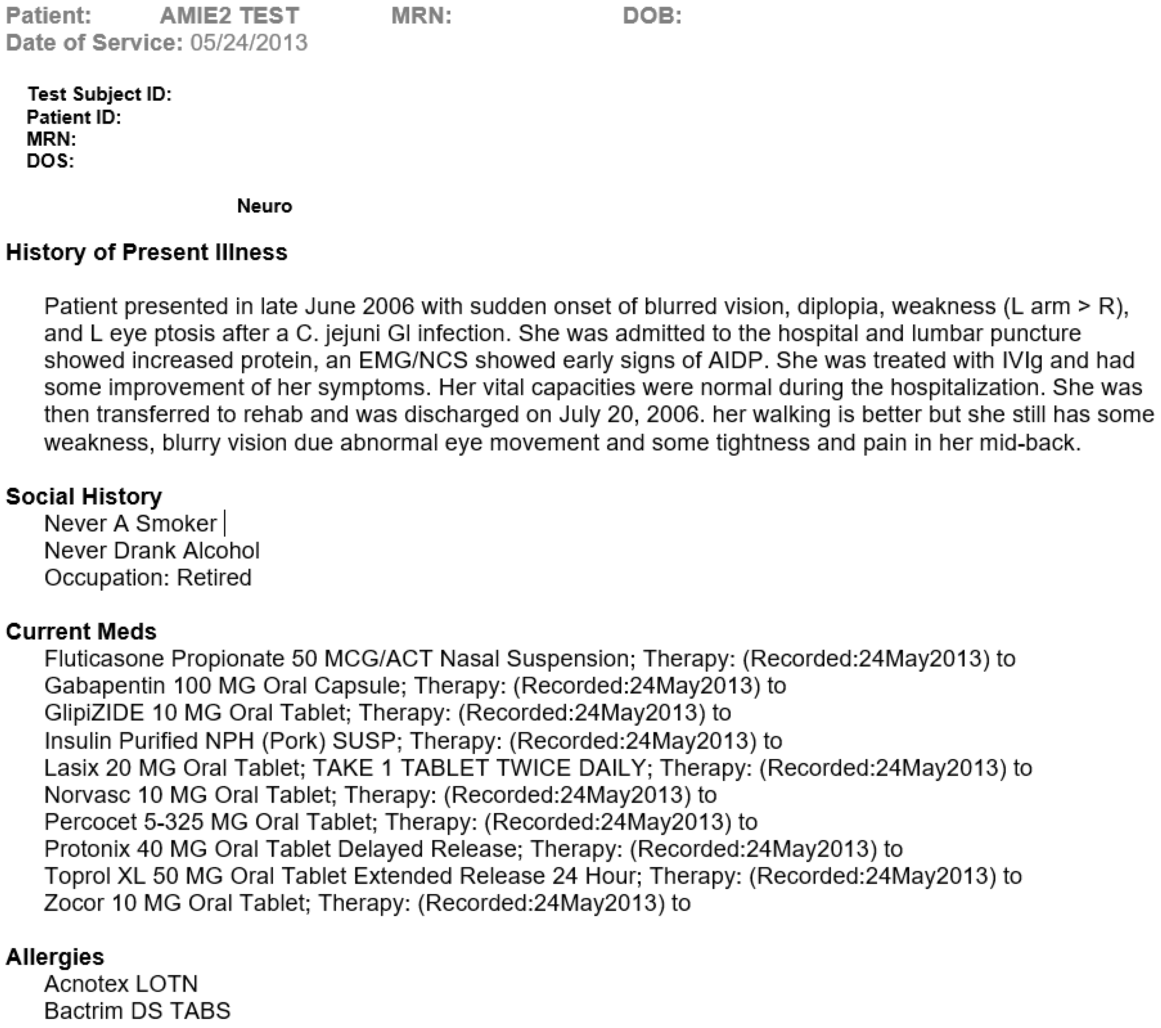
CURRENT MEDS:
Premarin 0.625 mg p.o. q.o.d., Aciphex 20 mg p.o. q. daily, Metoprolol 50 mg p.o. q. daily, Norvasc 5 mg p.o. q. daily, multivitamin, , B-complex vitamins and vitamin C daily.

PHYSICAL EXAM:
On examination today, this is a pleasant and healthy appearing woman. Blood pressure 154/72, heart rate 87, and weight 153 pounds. Pain is 0/10. Head is normocephalic and atraumatic. Head circumference is 54 cm, which is in the 10-25th percentile for a woman who is 5 foot and 6 inches tall. Spine is straight and nontender. Spinous processes are easily palpable. She has very mild kyphosis, but no scoliosis. There are no neurocutaneous stigmata. Regular rate and rhythm. No carotid bruits. No edema. No murmur. Peripheral pulses are good. Lungs are clear. Assessed for recent and remote memory, attention span, concentration, and fund of knowledge. She scored 30/30 on the MMSE when attention was tested with either spelling or calculations. She had no difficulty with visual structures. Pupils are equal. Extraocular movements are intact without nystagmus. Face is symmetric. Tongue and palate are midline. Jaw muscles strong. Cough is normal. SCM and shrug 5 and 5. Visual fields intact. Dix-Halpike maneuver did not produce nystagmus or vertigo. Normal for bulk, strength, and tone. There was no drift or tremor. Intact for pinprick and proprioception. Normal for finger-to-nose and no disdiadokokinesis. Are 2+ throughout. No Myerson. Assessed using the Tinetti assessment tool. She was fairly quick, but had some unsteadiness and a widened base. She did not need an assistive device. I gave her a score of 13/16 for balance and 9/12 for gait for a total score of 22/28. No postural instability.

DIAGNOSTIC RESULTS:
MRI was reviewed from June 26, 2008. It shows mild ventriculomegaly with a trace expansion into the temporal horns. The frontal horn span at the level of foramen of Munro is 3.8 cm with a flat 3rd ventricular contour and a 3rd ventricular span of 11 mm. The sylvian aqueduct is patent. There is no pulsation artifact. Her corpus callosum is bowed and effaced. She has a couple of small T2 signal abnormalities, but no significant periventricular signal change. There is transependymal absorption of

XSL•FO
RenderX

Figure 4. Example of part of the history and physical examination section of a neurology consultation note generated using the Standard-NLP protocol, illustrating the part of the note that was generated by Standard Entry. NLP: natural language processing.



Patient: **AMIE2 TEST** **MRN:** **DOB:**
Date of Service: 05/24/2013

Test Subject ID:
Patient ID:
MRN:
DOS:

Neuro

History of Present Illness

Patient presented in late June 2006 with sudden onset of blurred vision, diplopia, weakness (L arm > R), and L eye ptosis after a C. jejuni GI infection. She was admitted to the hospital and lumbar puncture showed increased protein, an EMG/NCS showed early signs of AIDP. She was treated with IVIg and had some improvement of her symptoms. Her vital capacities were normal during the hospitalization. She was then transferred to rehab and was discharged on July 20, 2006. her walking is better but she still has some weakness, blurry vision due abnormal eye movement and some tightness and pain in her mid-back.

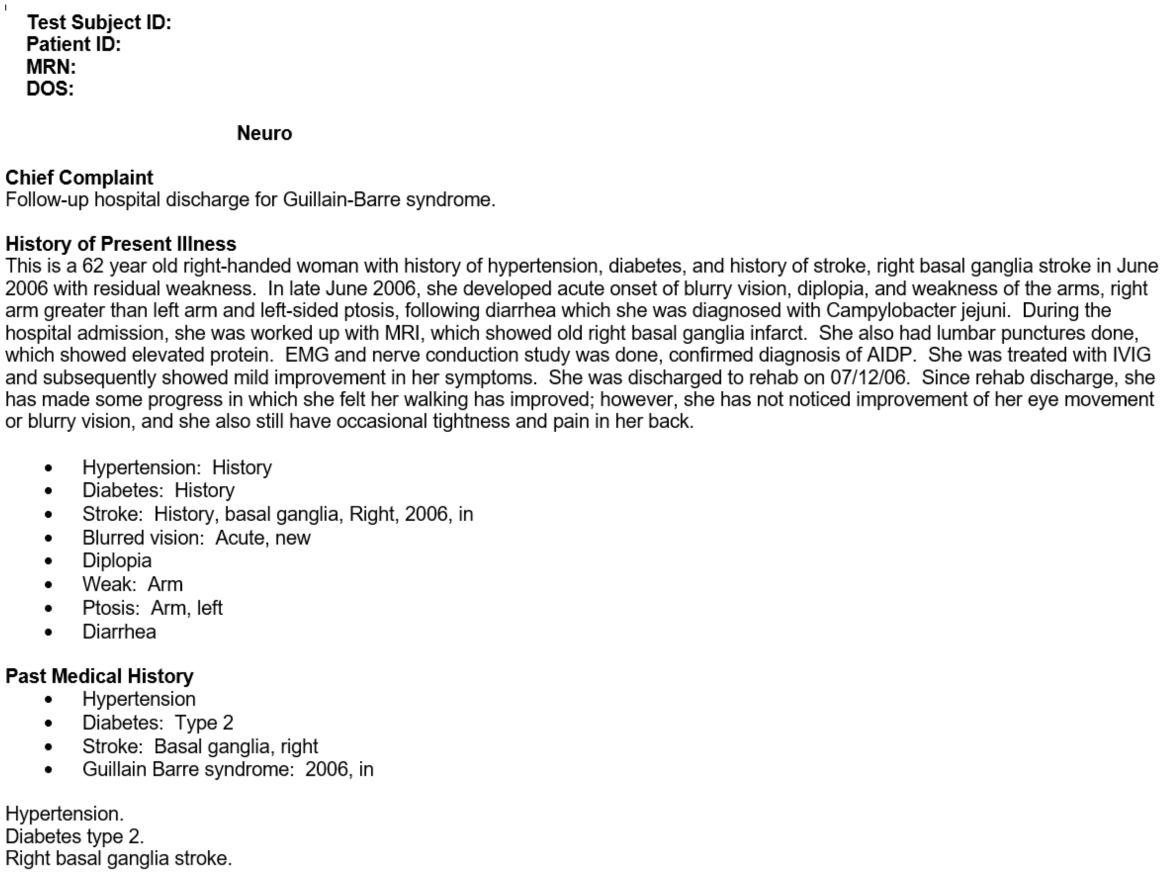
Social History
Never A Smoker
Never Drank Alcohol
Occupation: Retired

Current Meds
Fluticasone Propionate 50 MCG/ACT Nasal Suspension; Therapy: (Recorded:24May2013) to
Gabapentin 100 MG Oral Capsule; Therapy: (Recorded:24May2013) to
GlipiZIDE 10 MG Oral Tablet; Therapy: (Recorded:24May2013) to
Insulin Purified NPH (Pork) SUSP; Therapy: (Recorded:24May2013) to
Lasix 20 MG Oral Tablet; TAKE 1 TABLET TWICE DAILY; Therapy: (Recorded:24May2013) to
Norvasc 10 MG Oral Tablet; Therapy: (Recorded:24May2013) to
Percocet 5-325 MG Oral Tablet; Therapy: (Recorded:24May2013) to
Protonix 40 MG Oral Tablet Delayed Release; Therapy: (Recorded:24May2013) to
Toprol XL 50 MG Oral Tablet Extended Release 24 Hour; Therapy: (Recorded:24May2013) to
Zocor 10 MG Oral Tablet; Therapy: (Recorded:24May2013) to

Allergies
Acnotex LOTN
Bactrim DS TABS

XSL•FO
RenderX

Figure 5. Example of part of the history and physical examination section of a neurology consultation note generated using the NLP-Standard protocol, illustrating the part of the note that was generated by NLP Entry. NLP: natural language processing.



Test Subject ID:
Patient ID:
MRN:
DOS:

Neuro

Chief Complaint
Follow-up hospital discharge for Guillain-Barre syndrome.

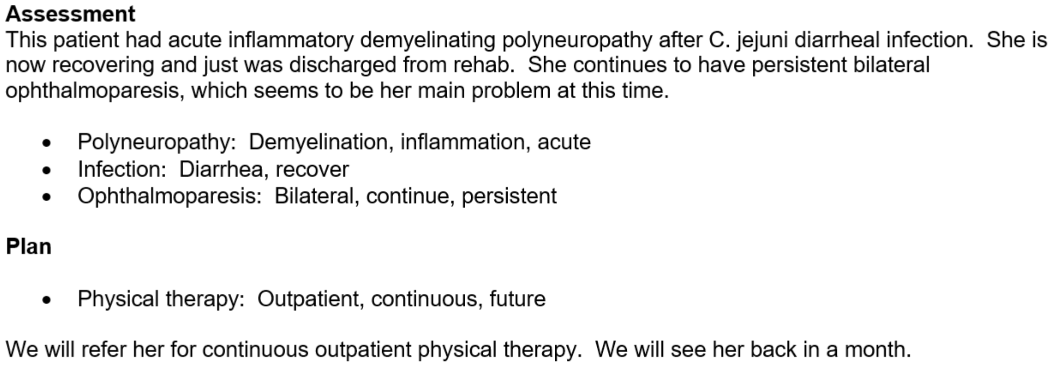
History of Present Illness
This is a 62 year old right-handed woman with history of hypertension, diabetes, and history of stroke, right basal ganglia stroke in June 2006 with residual weakness. In late June 2006, she developed acute onset of blurry vision, diplopia, and weakness of the arms, right arm greater than left arm and left-sided ptosis, following diarrhea which she was diagnosed with Campylobacter jejuni. During the hospital admission, she was worked up with MRI, which showed old right basal ganglia infarct. She also had lumbar punctures done, which showed elevated protein. EMG and nerve conduction study was done, confirmed diagnosis of AIDP. She was treated with IVIG and subsequently showed mild improvement in her symptoms. She was discharged to rehab on 07/12/06. Since rehab discharge, she has made some progress in which she felt her walking has improved; however, she has not noticed improvement of her eye movement or blurry vision, and she also still have occasional tightness and pain in her back.

- Hypertension: History
- Diabetes: History
- Stroke: History, basal ganglia, Right, 2006, in
- Blurred vision: Acute, new
- Diplopia
- Weak: Arm
- Ptosis: Arm, left
- Diarrhea

Past Medical History
- Hypertension
- Diabetes: Type 2
- Stroke: Basal ganglia, right
- Guillain Barre syndrome: 2006, in

Hypertension.
Diabetes type 2.
Right basal ganglia stroke.

Figure 6. Example of part of the assessment and plan section of a neurology consultation note generated using the Standard-NLP protocol, illustrating the part of the note that was generated by NLP Entry. NLP: natural language processing.



Assessment
This patient had acute inflammatory demyelinating polyneuropathy after C. jejuni diarrheal infection. She is now recovering and just was discharged from rehab. She continues to have persistent bilateral ophthalmoparesis, which seems to be her main problem at this time.

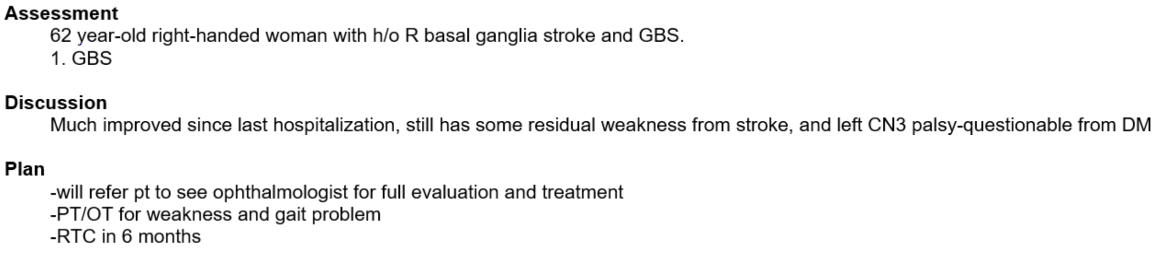
- Polyneuropathy: Demyelination, inflammation, acute
- Infection: Diarrhea, recover
- Ophthalmoparesis: Bilateral, continue, persistent

Plan

- Physical therapy: Outpatient, continuous, future

We will refer her for continuous outpatient physical therapy. We will see her back in a month.

Figure 7. Example of part of the assessment and plan section of a neurology consultation note generated using the NLP-Standard protocol, illustrating the part of the note that was generated by Standard Entry. NLP: natural language processing.



Assessment
62 year-old right-handed woman with h/o R basal ganglia stroke and GBS.
1. GBS

Discussion
Much improved since last hospitalization, still has some residual weakness from stroke, and left CN3 palsy-questionable from DM

Plan
-will refer pt to see ophthalmologist for full evaluation and treatment
-PT/OT for weakness and gait problem
-RTC in 6 months

XSL•FO
RenderX

Measures of Analysis

Time, Documentation Quality, and Usability

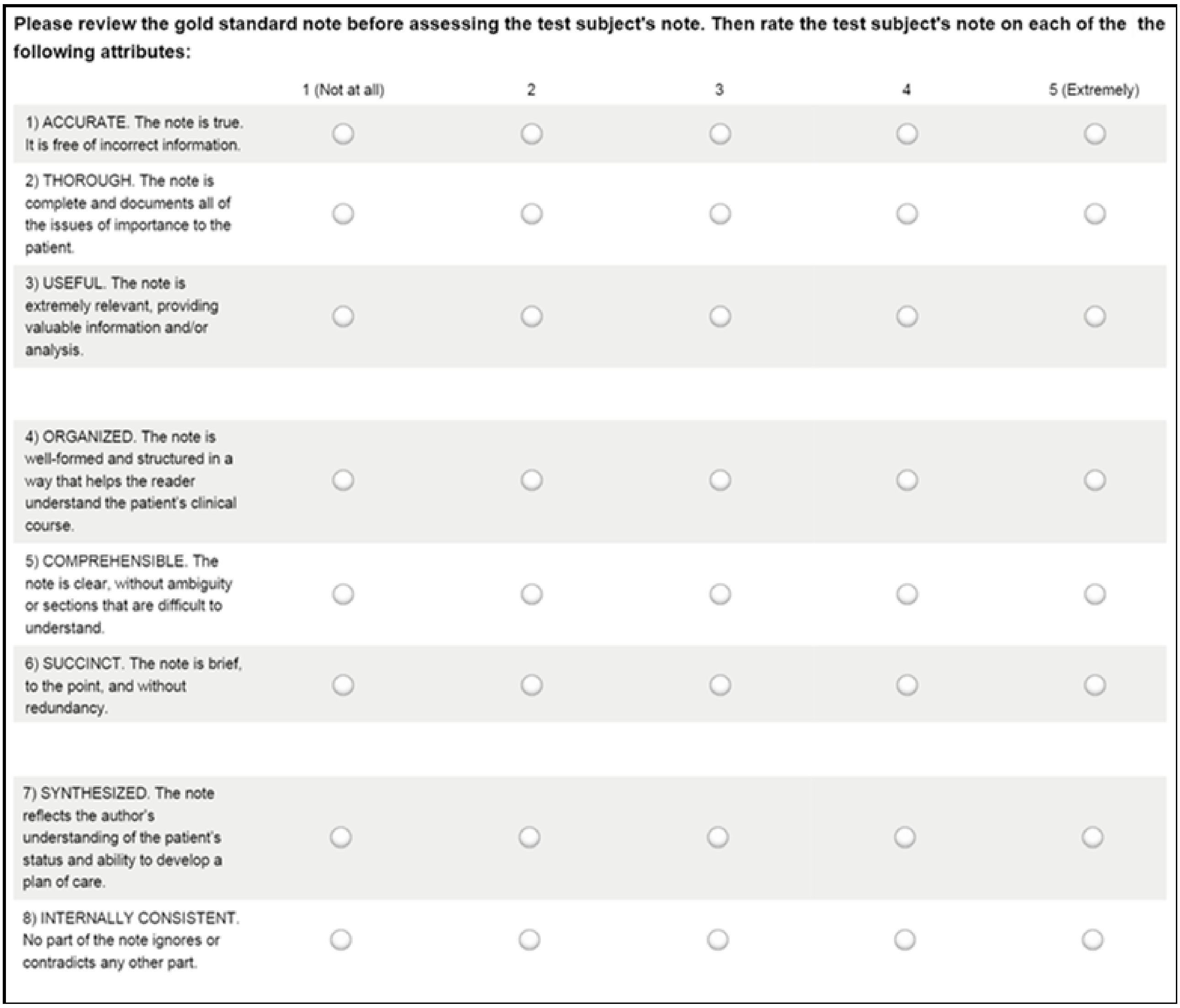
The time required for documentation was measured using a stopwatch.

The 4 expert clinicians (2 fellows and 2 attending physicians) were not participants and were blind to the protocols used to generate the documentation they evaluated. They were provided with gold standard versions of the test documentation they were asked to evaluate and told that the gold standard versions represented “high quality notes.” They were then instructed to measure documentation quality by comparing participants’ final test documentation against the gold standard versions of that documentation using the PDQI-9 tool [21], shown in Figure 8. The expert clinicians were given minimal background on the purpose for the study. They were independent and highly trained in their specialties (and they only graded documents within their own domains). In addition, they were provided with clear instructions and had a sound understanding of the PDQI-9 tool, which is known to be a reliable instrument [21]. Unfortunately, it was not practical to test interrater reliability. We used only 8 of the 9 PDQI-9 prompts (items) because one of the prompts required a judgment of whether the documentation was up-to-date and this was not meaningful in this particular context.

The gold standard versions of the documentation were generated by the expert clinicians in Microsoft Word. They produced these documents from clinical notes and modified them so that they were consistent with the clinical profile of the patient (ie, the patient’s assessment and treatment were consistent and derivable from history and physical examination findings). The expert clinicians were instructed to ensure that all elements of the documents were internally consistent and that they truly reflected a gold standard. The expert clinicians were compensated at a rate of US $125 per hour for their efforts. We did not have access to the interim work product of the expert clinicians. We were only provided with the expert clinicians’ grades.

The usability of the documentation processes was assessed using a modified version of the SUS [38]. Each participant completed 1 SUS questionnaire for each of the NLP-NLP and Standard-Standard protocols.

Figure 8. Physician Documentation Quality Instrument (PDQI-9) tool.



Please review the gold standard note before assessing the test subject's note. Then rate the test subject's note on each of the the following attributes:

	1 (Not at all)	2	3	4	5 (Extremely)
1) ACCURATE. The note is true. It is free of incorrect information.	○	○	○	○	○
2) THOROUGH. The note is complete and documents all of the issues of importance to the patient.	○	○	○	○	○
3) USEFUL. The note is extremely relevant, providing valuable information and/or analysis.	○	○	○	○	○
4) ORGANIZED. The note is well-formed and structured in a way that helps the reader understand the patient's clinical course.	○	○	○	○	○
5) COMPREHENSIBLE. The note is clear, without ambiguity or sections that are difficult to understand.	○	○	○	○	○
6) SUCCINCT. The note is brief, to the point, and without redundancy.	○	○	○	○	○
7) SYNTHESIZED. The note reflects the author's understanding of the patient's status and ability to develop a plan of care.	○	○	○	○	○
8) INTERNALLY CONSISTENT. No part of the note ignores or contradicts any other part.	○	○	○	○	○

XSL•FO
RenderX

Statistical Analysis

Data were analyzed using Intercooled Stata version 9.2 (StataCorp LP). Demographics were tabulated in regard to participants' years of EHR experience, years of experience with dictation, the number of cases per subject area, the frequency of use of each of the 4 protocols, dictation time, usability scores, and quality scores. The Shapiro-Wilk W test was used to determine whether continuous variables were normally distributed. Results are presented as mean (SD) or analysis of variance (ANOVA) results, median (interquartile range), or percentage; respectively, comparisons were made using *t* test, Wilcoxon rank sum analysis, or chi-square analysis.

Pearson correlation was performed on continuous variables. The association of years of EHR experience, years of experience with dictation, the 4 protocols (Standard-Standard, Standard-NLP, NLP-Standard, and NLP-NLP), and the 3 subject areas (cardiology, nephrology, and neurology) with the note dictation time was assessed using ANOVA.

Statistical significance was defined as alpha=.05 and Bonferroni correction was used where applicable for multiple comparisons.

Human Subjects Protection

The study was approved by the Institutional Review Board at Columbia University (#AAAK2458). All participants gave consent before their participation and were fully briefed on the true objectives of the study. The study protocol adhered to strict standards of confidentiality and privacy.

Results

A total of 31 unique individuals documented 3.8 (SD 0.7) notes on average. Of these, 28 participants completed all 4 protocols, 2 participants completed 2 protocols, and 1 participant completed 1 protocol. The participants who did not complete all 4 protocols were called away from the study and therefore could not finish the task. These individuals had an average EHR experience of 6.6 (SD 3.4) years (data were available for 30 individuals) and an average dictation experience of 2.8 (SD 5.6) years (data were available for 29 individuals). There was a significant association between years of EHR experience and years of dictation experience (*r*=.47, *P*=.01).

A total of 118 notes were documented across the 3 subject areas of cardiology (22/118, 18.6%), nephrology (21/118, 17.8%), and neurology (75/118, 63.6%). The Standard-Standard, Standard-NLP, NLP-Standard, and NLP-NLP protocols were used in 28/118 (23.7%), 28/118 (23.7%), 30/118 (25.4%), and 32/118 (27.1%) documented notes, respectively. The frequency of use of the 4 protocols was balanced across the 3 subject areas (see Table 3).

Table 3. Frequency of use of the 4 protocols by subject area for each documented note.

Protocol	Documented cardiology notes, n (%)	Documented nephrology notes, n (%)	Documented neurology notes, n (%)	Total number of documented notes, n (%)
Standard-Standard	5 (23)	5 (24)	18 (24)	28 (23.7)
Standard-NLP[a]	5 (23)	4 (19)	19 (25)	28 (23.7)
NLP-Standard	6 (27)	5 (24)	19 (25)	30 (25.4)
NLP-NLP	6 (27)	7 (33)	19 (25)	32 (27.1)
Total	22	21	75	118

[a] NLP: natural language processing.

Table 4. Median documentation time in minutes, with interquartile ranges, by protocol and subject area.

Protocol	Median (IQR[a]) time to document cardiology note (minutes)	Median (IQR) time to document nephrology note (minutes)	Median (IQR) time to document neurology note (minutes)
Standard-Standard	16.9 (16.5-19.7)	20.7 (18.6-23.2)	21.2 (17.6-29.9)
Standard-NLP[b]	13.8 (13.0-17.2)	21.3 (14.5-29.8)	18.7 (16.0-22.9)
NLP-Standard	7.5 (7.1-9.1)	12.1 (10.7-12.2)	11.0 (8.5-14.6)
NLP-NLP	5.2 (4.7-8.0)	7.3 (6.6-9.1)	8.5 (6.4-11.4)

[a] IQR: interquartile range.

[b] NLP: natural language processing.

XSL•FO
RenderX

Table 5. Interprotocol comparisons (Wilcoxon rank sum analysis).

Interprotocol comparisons	Statistical analysis of time difference (*P* value[a])		
	Cardiology notes	Nephrology notes	Neurology notes
Standard-Standard vs Standard-NLP[b]	.60	.81	.20
Standard-Standard vs NLP-Standard	.01	.03	<.001
Standard-Standard vs NLP-NLP	.006	.005	<.001
Standard-NLP vs NLP-Standard	.006	.05	.001
Standard-NLP vs NLP-NLP	.006	.008	<.001
NLP-Standard vs NLP-NLP	.11	.02	.02

[a] Statistical significance level: alpha=.0083 after Bonferroni correction.

[b] NLP: natural language processing.

Table 6. Document quality for each protocol (median values are presented).

Protocols and statistical comparisons	Document quality metrics[a]								
	A	T	U	O	C	S	Sy	I	Sum
Protocol, median score									
Standard-Standard (n=24)	3.5	3	4	4	4	4	4	4	29
Standard-NLP[b] (n=24)	4	4	4	3.5	4	2.5	4	4	29.5
NLP-Standard (n=27)	4	3	3	3	3	3	4	4	26
NLP-NLP (n=30)	4	4	3	3	3	2	3	4	24.5
Interprotocol comparisons, *P* value[c]									
Standard-Standard vs Standard-NLP		.04		.03		<.001			
Standard-Standard vs NLP-Standard					.04	.006			
Standard-Standard vs NLP-NLP				.002	.02	<.001	.03		
Standard-NLP vs NLP-Standard						.005			
Standard-NLP vs NLP-NLP							.02		
NLP-Standard vs NLP-NLP				.03		.001			

[a] The 8 document quality metrics are as follows: Accurate, Thorough, Useful, Organized, Comprehensible, Succinct, Synthesized, and Internally Consistent.

[b] NLP: natural language processing.

[c] Statistical significance level: alpha=.0083 after Bonferroni correction.

The documentation times were not normally distributed (z=4.6); thus, comparison of documentation times was performed using Wilcoxon rank sum analysis of medians. Table 4 summarizes the median time and interquartile range for the documentation of each note by protocol and subject area. Table 5 presents statistical analysis of interprotocol times. In each subject area, the NLP-NLP protocol required significantly less documentation time compared with either the Standard-Standard or Standard-NLP protocol. Compared with the Standard-Standard protocol, the NLP-Standard protocol required significantly less documentation time for the neurology subject area. Compared with the Standard-NLP protocol, the NLP-Standard protocol required significantly less documentation time for the cardiology and neurology subject areas.

On the basis of the ANOVA of documentation time, the model was statistically significant (adjusted R^2=.54, P<.001). This indicates that, taken together, the input variables used in the ANOVA model (EHR experience; years of experience with dictation; the 4 protocols, Standard-Standard, Standard-NLP, NLP-Standard, and NLP-NLP; and the 3 subject areas, cardiology, nephrology, and neurology) accounted for 54% of the variance in documentation time, the outcome variable. The factors significantly associated with documentation time included the protocol method (P<.001), subject area (P=.009), and the number of years of EHR experience (P=.047) but not the number of years of dictation experience (P=.77).

Document quality was assessed using 8 observed PDQI-9 metrics (Figure 8). Median values of the document quality metrics are presented for each protocol, in Table 6. Statistical

XSL•FO
RenderX

comparisons across the protocols are also presented in Table 6. The significant differences among the protocols occurred within the “Organized” metric (Standard-Standard vs NLP-NLP, 4 vs 3, *P*=.002) and the “Succinct” metric (Standard-Standard vs Standard-NLP, 4 vs 2.5, *P*<.001; Standard-Standard vs NLP-Standard, 4 vs 3, *P*=.006; Standard-Standard vs NLP-NLP, 4 vs 2, *P*<.001; Standard-NLP vs NLP-Standard, 2.5 vs 3, *P*=.005; and NLP-Standard vs NLP-NLP, 3 vs 2, *P*=.001).

The usability data were analyzed using a paired *t* test in a subset of 23 cases (n=5 cardiology, n=5 nephrology, and n=13 neurology) in which the same participant documented a case with both Standard-Standard and NLP-NLP protocols. The average duration of EHR experience for these 23 individuals was 6.3 (SD 2.8) years. The total score of the 10-component SUS measure was significantly higher when the participants used the NLP-NLP protocol compared with when they used the Standard-Standard protocol (mean 36.7, SD 5.4, compared with mean 30.3, SD 7.7; *P*=.007). Table 2 summarizes the usability scores and paired comparisons between the 2 protocols. Responses to 4 of the 10 usability questions (complexity, ease of use, learning the method very quickly, and cumbersomeness or awkwardness of use) significantly favored the NLP-NLP protocol over the Standard-Standard protocol.

Discussion

Findings

This formative study sought to assess the feasibility of using an EHR documentation method based on dictation and NLP by evaluating the effect of the method on documentation time, documentation quality, and usability. We found that a pure protocol of NLP Entry as well as hybrid protocols (involving both NLP Entry and Standard Entry) showed promise for EHR documentation, relative to Standard Entry alone (Standard-Standard Entry). It is our opinion that different parts of the note should be documented differently, but reaching a conclusion on the optimal method of documentation for each part of the note will require further study.

The finding that NLP-NLP Entry and NLP-Standard Entry required significantly less time than Standard-Standard Entry can be explained by the faster speed of dictation relative to that of entering data using the keyboard and mouse, rather than by the involvement of NLP.

No statistically significant difference was found between the overall documentation quality (measured using the PDQI-9 tool) of Standard-Standard Entry and that of any of the other 3 documentation protocols. The succinctness of Standard-Standard Entry documentation was found to be significantly greater than that of the other 3 protocols. This suggests that the note was judged to be more to the point and with less redundancy. In addition, documentation from Standard-Standard Entry was found to be more organized than that from NLP-NLP Entry, indicating that it was structured in a way that the reader could better understand the patient’s clinical course. When the participant used the Standard-Standard protocol, they used Standard Entry for history and physical examination sections as well as assessment and plan sections. When they used the Standard-NLP protocol, they used Standard Entry for history and physical examination sections and NLP Entry for assessment and plan sections. In the former (Standard-Standard), the participants tended to type shorter paragraphs for the assessment and plan sections. In the latter (Standard-NLP), they dictated the assessment and plan resulting in a larger volume of text. This difference warrants future scrutiny. On the basis of the results of the modified SUS, the participants’ usability ratings for NLP-NLP Entry were significantly higher than for Standard-Standard Entry. These findings suggest that, pending further study, EHR documentation methods using a combination of dictation and NLP show potential for reducing documentation time and increasing usability while maintaining documentation quality, relative to EHR documentation via standard keyboard-and-mouse entry.

Documentation methods using dictation and NLP have the potential to reduce some of the most egregious “pain points” for EHR data entry. These methods can facilitate capture and insertion of both structured data and transcribed text into the appropriate EHR sections, affording the user of the note the option of using one or both types of information. The structured data are ideal for interoperability and coding and may prove to be useful for analytics.

Opinion is divided regarding the relative advantages of narrative fields and structured fields in clinical documentation and in which contexts each excels or is preferable [39,40]. The flexibility to allow providers to enter text in narrative fields clashes with the desire to produce structured data to facilitate reuse of this information in EHRs [22]. It is this quest for achieving a balance of expressivity, richness, and completeness of detail in the patient health story, with reusable, and thus actionable, structured data in the patient EHR that motivates this field of inquiry. Improving the quality of structured notes generated within an EHR from the structured output of NLP Entry and transcribed text is an area requiring further study and development work.

Future Research

In future research, for the purpose of achieving documentation quality using dictation and NLP that, in all respects, is comparable to or better than documentation quality resulting from Standard Entry, certain changes to the NLP Entry process will be evaluated. We will assess the effects of requiring participants to use dictation under the constraints of a structured template on improving the organization, comprehensibility, succinctness, and synthesis of notes produced from NLP Entry. The templates would reflect the structure of the participant’s EHR. In addition, we will aim to improve the procedures by which NLP output data are translated into and transferred to the clinical note. We also plan to more systematically scrutinize data capture differences pertaining to documenting in different sections of the EHR note. This will enable us to fine-tune hybrid methods of data entry.

In a subsequent study, we will measure the time required for, and documentation quality and usability of, NLP Entry in live clinical use. This will require developing automated interfaces for sending the participants’ dictation to MediSapien NLP and for sending structured data and free text from MediSapien NLP

XSL•FO
RenderX

into the EHR, during which process the participant will be able to modify the documentation. This process will be facilitated by the emergence and widespread adoption of interoperability standards and messages that can carry rich structured data.

Limitations

This study has several limitations. One limitation is that the simulated interface used in this controlled experiment is somewhat lacking in ecological validity. In a real-world live setting, the structured data and the transcribed text data would both be inserted into the EHR via an automated interface. In addition, the physician would be able to review or modify the documentation before it was finalized. For the purposes of this formative study, this process was simplified. Therefore, an interface to automatically insert the structured data and text into the EHR or allow the physician to review the documentation before finalization was not used for this study. Instead, the insertion process was simulated by manually generating a note in Microsoft Word resembling one that might have been generated by the automated insertion process. Time required for generating the note was not included in the study's time measurements. To ensure that the manually generated note could have been produced by an automated process, it was produced following strict predetermined rules and without any reliance on human discretion.

Second, physicians generated documentation for the study based on test scripts about fictitious patient encounters. Test scripts included history and physical examination sections and were formatted as transcription notes. The assessment and treatment plan sections were excluded from the test scripts. Participants were instructed not to dictate or type verbatim what was written in the test scripts, but to understand what was written and document it in their own way. In addition to being instructed to generate history and physical examination sections, they were instructed to generate their own assessment and treatment plan sections, because those sections were excluded from the test script.

The sample size for cardiology and nephrology was rather small owing to recruiting challenges. This affected the power for determining differences for related contrasts. Clearly, a larger sample size would have enabled us to detect more subtle group differences.

A limitation of this method of generating test documentation was that because it presented medical information in a free-text format, it may have favored documentation methods requiring the physician to generate free text. NLP Entry requires documentation via dictation exclusively, and Standard Entry entails only some documentation via typing, with the rest entered by pointing and clicking using a mouse. Consequently, the results for time required to complete documentation may be biased toward free text and therefore toward NLP Entry. Nevertheless, we perceive a value in measuring the temporal differences and think that such differences may be consequential in real-world use of this system.

Conclusions

Current standard methods of EHR documentation have been shown to be extremely time consuming and are judged to have suboptimal usability. In this formative study, the feasibility of an approach to EHR data capture involving applying NLP to transcribed dictation was demonstrated. This approach was shown to have the potential to reduce the time required for documentation and improve usability while maintaining documentation quality in several respects. Future research will evaluate the NLP-based EHR data capture approach in a live clinical setting where generated structured data and transcribed text for real patients are inserted into the EHR via an automated interface.

The past decade has witnessed a dramatic increase in the adoption of EHRs as central instruments in medical practice. However, these systems have not yet proven to be reliable tools for facilitating clinical workflow or enhancing patient care. Recent advances in usability have led to the development of frameworks, new methods, and robust assessment tools that can be used to more precisely delineate the source of the problems associated with an interface [41,42]. In addition, novel approaches to design have provided new EHR approaches that better support flexibility and expressivity [43]. We anticipate that NLP-enabled data entry tools will form an important part of the solution space and will serve to enhance the user's experience.

There is ample evidence that clinicians spend many hours documenting patient records and sometimes at the expense of time that could be devoted to patient care. Dictation is a familiar method of data entry to most clinicians. The proposed solution leverages that familiarity and has the potential to produce a quality document or patient note in less time along with highly structured machine-readable codes.

Acknowledgments

The authors thank the clinicians who participated in the study. The research was supported by the Center for Advanced Information Management of New York State (#13240001) and the National Library of Medicine of the National Institutes of Health under Award Number R43LM011165. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of Interest

JM, AB, and AF are employed by ZyDoc Medical Transcription LLC, whose MediSapien NLP software product was used in this study.

XSL•FO
RenderX

References

1. Healthit. 2012. EHR Adoption & Meaningful Use Progress is Assessed in New Data Briefs URL: http://www.healthit.gov/buzz-blog/meaningful-use/ehr-adoption-meaningful-use-progress-assessed-data-briefs/ [accessed 2015-08-23] [WebCite Cache ID 6b0B1IIih]
2. Fiks AG, Alessandrini EA, Forrest CB, Khan S, Localio AR, Gerber A. Electronic medical record use in pediatric primary care. J Am Med Inform Assoc 2011;18(1):38-44 [FREE Full text] [doi: 10.1136/jamia.2010.004135] [Medline: 21134975]
3. Füchtbauer LM, Nørgaard B, Mogensen CB. Emergency department physicians spend only 25% of their working time on direct patient care. Dan Med J 2013 Jan;60(1):A4558. [Medline: 23340186]
4. McDonald C, Tang P, Hripcsak G. Electronic health record systems. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics. London: Springer; 2014:255-284.
5. Cimino JJ. Improving the electronic health record–are clinicians getting what they wished for? JAMA 2013 Mar 13;309(10):991-992 [FREE Full text] [doi: 10.1001/jama.2013.890] [Medline: 23483171]
6. Hsiao C, Hing E. 2014. Use and Characteristics of Electronic Health Record Systems Among Office-based Physician Practices: United States, 2001–2013. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. Atlanta URL: http://www.cdc.gov/nchs/data/databriefs/db143.htm [accessed 2015-08-23] [WebCite Cache ID 6b0OxO6Wy]
7. Slight SP, Berner ES, Galanter W, Huff S, Lambert BL, Lannon C, et al. Meaningful use of electronic health records: experiences from the field and future opportunities. JMIR Med Inform 2015;3(3):e30 [FREE Full text] [doi: 10.2196/medinform.4457] [Medline: 26385598]
8. Friedberg M, Chen P, Van BK, Aunon F, Pham C, Caloyeras J, et al. Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. Santa Monica, CA: Rand Corporation; 2013. URL: http://www.rand.org/content/dam/rand/pubs/research_reports/RR400/RR439/RAND_RR439.pdf [WebCite Cache ID 6ds7RJ7In]
9. Middleton B, Bloomrosen M, Dente MA, Hashmat B, Koppel R, Overhage JM, et al. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. J Am Med Inform Assoc 2013 Jun;20(e1):e2-e8 [FREE Full text] [doi: 10.1136/amiajnl-2012-001458] [Medline: 23355463]
10. Westbrook JI, Baysari MT, Li L, Burke R, Richardson KL, Day RO. The safety of electronic prescribing: manifestations, mechanisms, and rates of system-related errors associated with two commercial systems in hospitals. J Am Med Inform Assoc 2013;20(6):1159-1167 [FREE Full text] [doi: 10.1136/amiajnl-2013-001745] [Medline: 23721982]
11. Schumacher R, Lowry S. Customized common industry format template for electronic health record usability testing. National Institute of Standards and Technology, US Department of Commerce. URL: http://www.nist.gov/healthcare/usability/upload/LowryNISTIR-7742Customized_CIF_Template_for_EHR_Usability_Testing_Publicationl_Version-doc.pdf [accessed 2015-12-10] [WebCite Cache ID 6dfsIiEmI]
12. Schumacher R, Lowry S. IST guide to the processes approach for improving the usability of electronic health records. National Institute of Standards and Technology, US Department of Commerce; 2010 Nov 1. URL: http://www.nist.gov/healthcare/usability/upload/Guide_Final_Publication_Version.pdf [WebCite Cache ID 6dfslYqDT]
13. Armijo D, McDonnell C, Werner K. Electronic health record usability: evaluation and use case framework. Healthit: Agency for Healthcare Research and Quality; 2009. URL: https://healthit.ahrq.gov/EHR_evaluation_and_use_case_framework [WebCite Cache ID 6dftDJ6zr]
14. Ohno-Machado L. Electronic health records and computer-based clinical decision support: are we there yet? J Am Med Inform Assoc 2011 Mar 01;18(2):109. [doi: 10.1136/amiajnl-2011-000141]
15. Payne TH, Corley S, Cullen TA, Gandhi TK, Harrington L, Kuperman GJ, et al. Report of the AMIA EHR-2020 Task Force on the status and future direction of EHRs. J Am Med Inform Assoc 2015 Sep;22(5):1102-1110. [doi: 10.1093/jamia/ocv066] [Medline: 26024883]
16. Kimia AA, Savova G, Landschaft A, Harper MB. An introduction to natural language processing: how you can get more from those electronic notes you are generating. Pediatr Emerg Care 2015 Jul;31(7):536-541. [doi: 10.1097/PEC.0000000000000484] [Medline: 26148107]
17. Bernat JL. Ethical and quality pitfalls in electronic health records. Neurology 2013 Mar 12;80(11):1057-1061. [doi: 10.1212/WNL.0b013e318287288c] [Medline: 23479465]
18. Cusack CM, Hripcsak G, Bloomrosen M, Rosenbloom ST, Weaver CA, Wright A, et al. The future state of clinical data capture and documentation: a report from AMIA's 2011 Policy Meeting. J Am Med Inform Assoc 2013 Jan 1;20(1):134-140 [FREE Full text] [doi: 10.1136/amiajnl-2012-001093] [Medline: 22962195]
19. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 1;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
20. Office of the National Coordinator of Health Information Technology. Capturing high quality electronic health records data to support performance improvement: a learning guide. 2013. URL: http://www.healthit.gov/sites/default/files/onc-beacon-lg3-ehr-data-quality-and-perform-impvt.pdf [accessed 2015-08-23] [WebCite Cache ID 6b0HF2qpV]

XSL•FO
RenderX

21. Stetson PD, Bakken S, Wrenn JO, Siegler EL. Assessing electronic note quality using the Physician Documentation Quality Instrument (PDQI-9). Appl Clin Inform 2012;3(2):164-174 [FREE Full text] [Medline: 22577483]
22. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc 2011;18(2):181-186 [FREE Full text] [doi: 10.1136/jamia.2010.007237] [Medline: 21233086]
23. Suominen H, Zhou L, Hanlen L, Ferraro G. Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations. JMIR Med Inform 2015;3(2):e19 [FREE Full text] [doi: 10.2196/medinform.4321] [Medline: 25917752]
24. Friedman C, Elhadad N. Natural language processing in health care and biomedicine. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics. 4th edition. London: Springer; 2014:255-284.
25. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 Jun;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
26. Wu S, Zhu D, Hersh W, Liu H. Clinical information retrieval with split-layer language models. Presented at: Proceedings of the Association for Computing Machinery SIGIR workshop on health search and discovery (HSD); August 1, 2013; Dublin, Ireland. URL: https://www.eecis.udel.edu/~zhu/uploads/hsd2013_paper.pdf
27. Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques.. Computational Linguistics 2007 Mar;33(1):63-103. [doi: 10.1162/coli.2007.33.1.63]
28. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc 2011 Oct;18(5):544-551 [FREE Full text] [doi: 10.1136/amiajnl-2011-000464] [Medline: 21846786]
29. Liu G, Weng C, Yu H. Natural language processing, electronic health records, and clinical research. In: Richesson RL, Andrews JR, editors. Clinical Research Informatics. London: Springer; 2012:293-310.
30. MModal. M*Modal's cloud-based artificial intelligence is widely adopted by physicians to improve patient care. Press release. URL: https://mmodal.com/resources/press/mmodals-cloud-based-artificial-intelligence-is-widely-adopted-by-physicians-to-improve-patient-care/ [accessed 2016-06-14] [WebCite Cache ID 6iGbdZbuh]
31. Nuance. Meaningful use: preserving the health story, providing structure for the EHR. URL: http://www.nuance.com/ucmprod/groups/healthcare/@web-enus/documents/collateral/nd_003313.pdf [accessed 2016-06-15] [WebCite Cache ID 6iHuujbOD]
32. Sun L. Sep 10. The rise of speech recognition technology in EHRs. The Motley Fool; 2013 Sep 10. URL: http://www.fool.com/investing/general/2013/09/10/a-closer-look-at-speech-recognition-technology-in.aspx [accessed 2016-10-23] [WebCite Cache ID 6b0KeiRV6]
33. Wagholikar KB, MacLaughlin KL, Henry MR, Greenes RA, Hankey RA, Liu H, et al. Clinical decision support with automated text processing for cervical cancer screening. J Am Med Inform Assoc 2012;19(5):833-839 [FREE Full text] [doi: 10.1136/amiajnl-2012-000820] [Medline: 22542812]
34. Kohn MS, Sun J, Knoop S, Shabo A, Carmeli B, Sow D, et al. IBM's Health Analytics and clinical decision support.. Yearb Med Inform 2014;9:154-162 [FREE Full text] [doi: 10.15265/IY-2014-0002] [Medline: 25123736]
35. Wu Y, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Xu H. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. AMIA Annu Symp Proc 2012;2012:997-1003 [FREE Full text] [Medline: 23304375]
36. Salmasian H, Freedberg DE, Friedman C. Deriving comorbidities from medical records using natural language processing. J Am Med Inform Assoc 2013 Dec;20(e2):e239-e242 [FREE Full text] [doi: 10.1136/amiajnl-2013-001889] [Medline: 24177145]
37. Brooke J. SUS: A quickdirty usability scale. In: Jordan PW, Thomas B, Lyall I, editors. Usability Evaluation in Industry. London: Taylor & Francis, Ltd; 1996:189-194.
38. US Department of Health & Human Services. System Usability Scale (SUS). URL: http://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html [accessed 2015-08-23] [WebCite Cache ID 6b0M4XSKc]
39. Jacobs L. Interview with Lawrence Weed, MD—the father of the problem-oriented medical record looks ahead.. Perm J 2009;13(3):84-89 [FREE Full text] [Medline: 20740095]
40. Johnson SB, Bakken S, Dine D, Hyun S, Mendonça E, Morrison F, et al. An electronic health record based on structured narrative. J Am Med Inform Assoc 2008;15(1):54-64 [FREE Full text] [doi: 10.1197/jamia.M2131] [Medline: 17947628]
41. Zhang J, Walji MF. TURF: toward a unified framework of EHR usability. J Biomed Inform 2011 Dec;44(6):1056-1067 [FREE Full text] [doi: 10.1016/j.jbi.2011.08.005] [Medline: 21867774]
42. Zheng K, Haftel HM, Hirschl RB, O'Reilly M, Hanauer DA. Quantifying the impact of health IT implementations on clinical workflow: a new methodological perspective. J Am Med Inform Assoc 2010;17(4):454-461 [FREE Full text] [doi: 10.1136/jamia.2010.004440] [Medline: 20595314]
43. Senathirajah Y, Bakken S, Kaufman D. The clinician in the driver's seat: part 1 – a drag/drop user-composable electronic health record platform. J Biomed Inform 2014 Dec;52:165-176 [FREE Full text] [doi: 10.1016/j.jbi.2014.09.002] [Medline: 25240253]

XSL•FO
RenderX

Abbreviations

ANOVA: analysis of variance
CBD: computer-based documentation
CUMC: Columbia University Medical Center
EHR: electronic health record
HITECH: Health Information Technology for Economic and Clinical Health
MedLEE: Medical Language Extraction and Encoding System
NLP: natural language processing
PDQI: Physician Documentation Quality Instrument
SUS: System Usability Scale

Edited by G Eysenbach; submitted 20.01.16; peer-reviewed by PY Yen, C Lovis; comments to author 22.02.16; revised version received 21.06.16; accepted 15.09.16; published 28.10.16.

Please cite as:
Kaufman DR, Sheehan B, Stetson P, Bhatt AR, Field AI, Patel C, Maisel JM
Natural Language Processing–Enabled and Conventional Data Capture Methods for Input to Electronic Health Records: A Comparative Usability Study
JMIR Med Inform 2016;4(4):e35
URL: http://medinform.jmir.org/2016/4/e35/
doi:10.2196/medinform.5544
PMID:27793791

©David R. Kaufman, Barbara Sheehan, Peter Stetson, Ashish R. Bhatt, Adele I. Field, Chirag Patel, James Mark Maisel. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 28.10.2016. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Finding Important Terms for Patients in Their Electronic Health Records: A Learning-to-Rank Approach Using Expert Annotations

Jinying Chen[1], PhD; Jiaping Zheng[2], MS; Hong Yu[1,3], PhD, FACMI

[1]Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, United States

[2]School of Computer Science, University of Massachusetts, Amherst, MA, United States

[3]Bedford Veterans Affairs Medical Center, Center for Healthcare Organization and Implementation Research, Bedford, MA, United States

Corresponding Author:
Jinying Chen, PhD
Department of Quantitative Health Sciences
University of Massachusetts Medical School
368 Plantation Street
Worcester, MA
United States
Phone: 1 774 455 3527
Fax: 1 508 856 8993
Email: jinying.chen@umassmed.edu

Abstract

Background: Many health organizations allow patients to access their own electronic health record (EHR) notes through online patient portals as a way to enhance patient-centered care. However, EHR notes are typically long and contain abundant medical jargon that can be difficult for patients to understand. In addition, many medical terms in patients’ notes are not directly related to their health care needs. One way to help patients better comprehend their own notes is to reduce information overload and help them focus on medical terms that matter most to them. Interventions can then be developed by giving them targeted education to improve their EHR comprehension and the quality of care.

Objective: We aimed to develop a supervised natural language processing (NLP) system called Finding impOrtant medical Concepts most Useful to patientS (FOCUS) that automatically identifies and ranks medical terms in EHR notes based on their importance to the patients.

Methods: First, we built an expert-annotated corpus. For each EHR note, 2 physicians independently identified medical terms important to the patient. Using the physicians’ agreement as the gold standard, we developed and evaluated FOCUS. FOCUS first identifies candidate terms from each EHR note using MetaMap and then ranks the terms using a support vector machine-based learn-to-rank algorithm. We explored rich learning features, including distributed word representation, Unified Medical Language System semantic type, topic features, and features derived from consumer health vocabulary. We compared FOCUS with 2 strong baseline NLP systems.

Results: Physicians annotated 90 EHR notes and identified a mean of 9 (SD 5) important terms per note. The Cohen’s kappa annotation agreement was .51. The 10-fold cross-validation results show that FOCUS achieved an area under the receiver operating characteristic curve (AUC-ROC) of 0.940 for ranking candidate terms from EHR notes to identify important terms. When including term identification, the performance of FOCUS for identifying important terms from EHR notes was 0.866 AUC-ROC. Both performance scores significantly exceeded the corresponding baseline system scores (*P*<.001). Rich learning features contributed to FOCUS’s performance substantially.

Conclusions: FOCUS can automatically rank terms from EHR notes based on their importance to patients. It may help develop future interventions that improve quality of care.

(JMIR Med Inform 2016;4(4):e40) doi:10.2196/medinform.6373

KEYWORDS

electronic health records; natural language processing; information extraction; supervised learning; learning to rank

XSL•FO
RenderX

Introduction

Background and Significance

Greater patient involvement is indispensable in delivering high-quality patient-centered care. In one effort to achieve this goal, spurred by the Health Information Technology for Economic and Clinical Health Act [1,2] and the Centers for Medicare and Medicaid Services Medicare Electronic Health Record (EHR) incentive program [3], online patient portals have been widely adopted by health systems in the United States [3,4]. In addition to giving patients structured information from EHRs (eg, laboratory test results and medication lists), the OpenNotes initiative [5] and the Blue Button movement [6] allow patients to access their full EHR notes through patient portals. Early evidence shows improved medical comprehension, health care management, and outcomes from the OpenNotes initiative [7-9].

However, the benefits from accessing their full EHR notes would be compromised if patients cannot comprehend their notes. EHRs were created for physician-physician communication, and thus are frequently long and contain abundant medical jargon. Patients who usually do not have the same medical training as physicians are likely overwhelmed by the medical jargon, and therefore face an enormous challenge in comprehending their notes. For example, EHRs were written at an 8th-12th-grade reading level [10-13], which is above the average adult patient's reading level of 7th-8thgrade in the United States [14-19]. In addition, 36% of adult Americans have limited health literacy [19] and have shown difficulty in comprehending medical jargon [20-25]. In fact, limited health literacy has been identified as one of the major barriers to patient online portal use, which includes the interpretation of information from EHRs [26-28]. Therefore, information technologies that support EHR comprehension are much needed to supplement the widespread use of patient portals and EHRs among patients.

To support patient EHR comprehension, this work focuses on identifying medical terms that matter most to individual patients in their EHR notes—we used the 2 phrases "medical terms" and "medical jargon" interchangeably in this paper. Our work was motivated by 2 reasons. First, medical terms, which are fundamental to discourse-level EHR comprehension, have been shown to be obstacles for patients [20-25]. Second, EHR notes incorporate a comprehensive description of patients' medical courses yet patients may care about their immediate concerns. For example, a radiology report may describe technical details of tumor images; however, the patient may want to know only the tumor size, the diagnosis, and the prognosis. When helping patients comprehend their own EHR notes, the approach of explaining all the jargon in their notes may likely overwhelm them and may be unnecessary in the first place.

Therefore, in this study we identify medical jargon most important to individual patients. Personalized interventions can then be developed by giving targeted educational materials to each individual patient.

In order to find out whether medical terms can be prioritized, we asked physicians to identify terms important to patients in EHRs. Textbox 1 shows an excerpt from a typical EHR note from our corpus. Although there are many medical terms in this piece of text—here we only highlighted a subset of terms identified by MetaMap [29] for illustration purposes—physicians identified only 5 terms most important for patients to know: *thrombocytosis*, *Crohn disease*, *budesonide*, *diabetes mellitus*, and *metformin*. Note that physicians do not mark many unfamiliar medical terms (eg, *complete blood count* [*CBC*], *hematemesis*, and *epistaxis*), suggesting that they do not rank terms based on their difficulty levels.

Textbox 1. A sample electronic health record text where physicians identified important medical terms (bracketed with angle brackets). Other medical terms are italicized.

> xxx is a xx-year-old man referred for evaluation of <thrombocytosis>. Prior *CBCs* from xxx through xxx revealed *platelet counts* ranging from 400,000 to 500,000, but no more recent studies are available. He has long-standing <Crohn disease> and although he says he has not had *gastrointestinal bleeding* in the past, he has been given iron, which he is taking twice daily. He has black stool, but notes no blood and he has not had *hematemesis*. He notes no blood in his urine or sputum and he has no *epistaxis*. He discontinued the use of iron yesterday because he thought that might alleviate his gastrointestinal complaints, but he does not feel different today. He is cared for by Dr. xxx at xxx Hospital Medical Center in xxx. He has no history of prior cancers, *tuberculosis* or other infectious diseases. He has been taking <budesonide> for his <Crohn disease>. He has no unexplained fevers, although he states he often feels hot. He has no soaking sweats and has not had unexplained weight loss. He believes he was referred to an *oncologist* many years ago at xxx, but he cannot recall the reason for that referral, who the doctor was, or what the findings were. He often feels queasy and nauseated, but has no vomiting. He has loose stools up to 4 days per week, but has had a stable pattern of <Crohn disease>. Also notable for <diabetes mellitus> for which he takes <metformin> and has required no *insulin* and has had no complications of *retinopathy* or *renal dysfunction*. <Crohn disease> as described above and an enlarged prostate.

Our aim was to develop a supervised natural language processing (NLP) system called Finding impOrtant medical Concepts most Useful to patientS (FOCUS) to automatically rank those EHR (patient)-specific important terms as high. This task was challenging, as the problem could not be solved by using only simple strategies such as term unfamiliarity, term frequency, and handcrafted rules (details in the Discussion section). We therefore built FOCUS with supervised learning and rich features.

To the best of our knowledge, our work is the first to successfully rank medical terms in EHR notes by focusing on patients' needs. This is an important step toward information reduction and personalized interventions to improve patient EHR comprehension. Our contributions are multifold. First, we defined a new NLP task of prioritizing or ranking medical terms that are important for patients. Second, we developed a state-of-the-art learning-based NLP system to automate the task. Third, we explored novel semantically motivated learning features.

XSL•FO
RenderX

By using a robust learning framework, FOCUS can be readily adapted to other NLP tasks including summarization and question answering.

Related Works

Natural Language Processing Systems Facilitating Concept-Level Electronic Health Record Comprehension

There has been active research on linking medical terms to lay terms [11,30,31], consumer-oriented definitions [12] and educational materials [32], and showing improved comprehension with such interventions [11,12].

On the issue of determining which medical terms to simplify, there is previous work that used frequency-based and/or context-based approaches to check if a term is unfamiliar to the average patient or if it has simpler synonyms [11,30,31]. Such work focuses on identifying difficult medical terms and treats these terms as equally important.

Our approach is different in 2 aspects: (1) we focus on finding important medical terms, which are not equivalent to difficult medical terms, as discussed in the Background and Significance subsection; and (2) our approach is patient centered and prioritizes important terms for each EHR note of individual patients. We developed several learning features, including term frequency, term position, term frequency-inverse document frequency (TF-IDF), and topic feature, to serve this purpose.

It is worth noting that our approach is complementary to previous work. For example, in a real-world application, we can display the lay definitions for all the difficult medical terms in a patient's EHR note, and then highlight those terms that FOCUS predicts to be most important to this patient.

Single-Document Keyphrase Extraction

Our work is inspired by, but different from, single-document keyphrase extraction (KE), which identifies terms or phrases representing important concepts and topics in a document. KE targets topics that the writers wanted to convey when writing the documents. Unlike KE, our work does not focus on topics important to physicians (ie, the writers and the target readers when writing the EHR notes), but rather focuses on patients, the new readers of the notes.

Both supervised and unsupervised methods have been developed for KE [33]. We use supervised methods, which in general perform better than unsupervised ones when training data is available.

Most supervised methods formulate KE as a binary classification problem. The confidence scores output by the classification algorithms are used to rank candidate phrases. Various algorithms have been explored, such as naïve Bayes, decision tree, bagging, support vector machine (SVM), multilayer perceptron, and random forest (RF) [34-43]. In our study, we implemented RF [43] as a strong baseline system.

KE in the biomedical domain mainly focused on literature articles and domain-specific methods and features [44-47]. For example, Li et al [44] developed a software tool called keyphrase identification program (KIP) to extract keyphrases from medical articles. KIP used Medical Subject Headings (MeSH) as the knowledge base to compute a score to reflect a phrase's domain specificity. It assigned each candidate phrase a rank score by multiplying its within-document term frequency and domain-specificity score.

Different from the aforementioned approaches, we treat KE as a ranking problem and use the ranking SVM (rankSVM) approach [48] as it has been shown to be effective in KE in scientific literature, news, and weblogs [42].

Common learning features used by previous work include frequency-based features (eg, TF-IDF), term-related features (eg, the term itself, its position in a document, and its length), document structure-based features (eg, whether a term occurs in the title or abstract of a scientific paper), and syntactic features (eg, the part-of-speech [POS] tags). Features derived from external resources, such as Wikipedia and query logs, have also been used to represent term importance [39,40]. Unlike previous work, we explored rich semantic features specifically available to the medical domain.

Medelyan and Witten [45] developed a system that extends the widely used keyphrase extraction algorithm KEA [34] by using semantic information from domain-specific thesauri, which they called KEA++. KEA++ has been applied to the medical domain, where it used MeSH vocabulary to extract candidate phrases from medical articles and used MeSH concept relations to compute its domain-specific feature. In this study, we adapted KEA++ to the EHR data and used the adapted KEA++ as a strong baseline system.

Methods

A FOCUS Corpus of Electronic Health Records With Expert-Annotated Important Concepts

We created a FOCUS corpus, which is a collection of 90 representative EHR discharge summaries and progress notes from the University of Massachusetts Memorial Hospital outpatient clinics. To maximize the representativeness, we selected notes from patients with 6 different but common primary clinical diagnoses: cancer, chronic obstructive pulmonary disease, diabetes, heart failure, hypertension, and liver failure. We deidentified the notes and then asked physicians to identify, for each note, terms important to patients.

We adopted the expert annotation approach for this study for the following reasons. First, annotating important medical terms requires full comprehension of an EHR note. Such level of comprehension may be beyond the capacity of average patients [11-13,30]. Previous work shows that even lay people with higher education (ie, college or graduate degrees) have difficulty with comprehending EHR notes [11,30]. Second, physicians have specific medical training for communicating with patients and understanding their needs. Physicians' expertise would guide patients in understanding the most important aspects that are medically relevant to their health and well-being.

We developed an annotation guideline (see Multimedia Appendix 1) to instruct physicians to identify at least 5 of the most important medical terms per EHR note, which the patients need to know in order to comprehend the note for the most

XSL•FO
RenderX

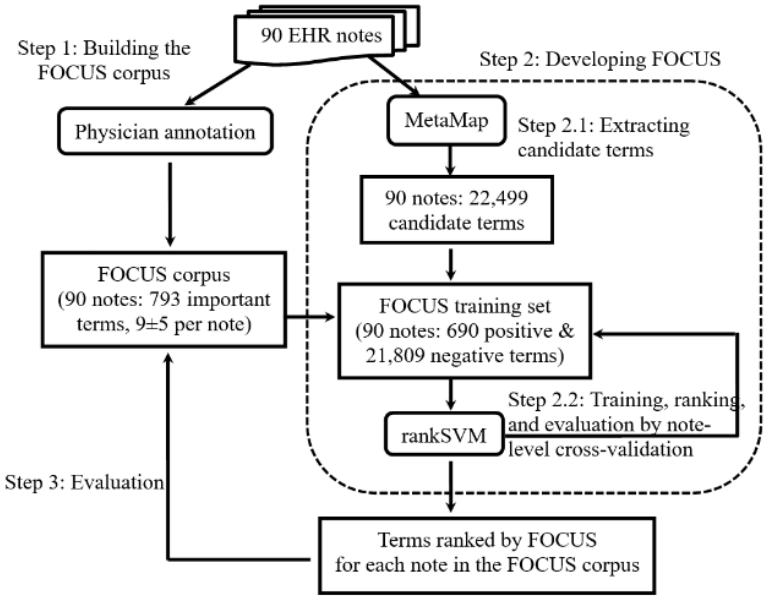
important aspects medically relevant to their health and treatment course. For each note, we obtained annotations from 2 physicians and used the agreement from both physicians as the gold standard for our experiments. Three physicians did the annotation and annotated 48, 68, and 64 notes, respectively.

FOCUS

Overview

Figure 1 shows the overview of FOCUS and its corpus and evaluation. In Step 2 of the approach, FOCUS first extracts candidate terms (Step 2.1) and then ranks them (Step 2.2). Since we focused on ranking in this study, we used MetaMap [29], a widely used medical concept detection tool, to automatically identify candidate terms from each EHR note. We then applied rankSVM to rank the terms.

Figure 1. Overview of our approach: building the FOCUS corpus (Step 1), developing FOCUS (Step 2), and evaluation (Step 3). FOCUS: Finding impOrtant medical Concepts most Useful to patientS; EHR: electronic health record; rankSVM: ranking support vector machine.


90 EHR notes
Step 1: Building the FOCUS corpus
Step 2: Developing FOCUS
Physician annotation
MetaMap
Step 2.1: Extracting candidate terms
90 notes: 22,499 candidate terms
FOCUS corpus (90 notes: 793 important terms, 9±5 per note)
FOCUS training set (90 notes: 690 positive & 21,809 negative terms)
Step 2.2: Training, ranking, and evaluation by note-level cross-validation
rankSVM
Step 3: Evaluation
Terms ranked by FOCUS for each note in the FOCUS corpus

Ranking Support Vector Machine

RankSVM [48] is a pairwise ranking method, which can learn to rank important terms in each EHR note as higher than nonimportant ones.

Our training data for rankSVM contain the following: (1) a set *E* of EHR notes; (2) a list of candidate terms T_e associated with each EHR note *e*; and (3) for a term *t* T_e, a *d*-dimension feature vector x_t $R^d$and a binary target value (ie, label) y_t which denotes whether *t* is an important medical term in *e*. In our case, y_t is 1 if *t* is important in *e* and 0 if not. In the general framework of ranking, y_t corresponds to the ranking order of *t*, and the more important *t* is, the higher order and the larger value of y_t it has.

Let *P* be the set of term pairs (*i*, *j*), where term *i* and term *j* occur in the same EHR note and term *i* is important (y_i=1) and term *j* is not important (y_j=0) (ie, P={ (i, j) | $y_i > y_j$}). The rankSVM model is built by minimizing the objective function [48], as defined by equation 1 in Figure 2, where *w* is the feature weight vector; $\varepsilon_{i,j}$ is the slack variable that measures the model's soft-margin error for term pair (*i*, *j*); *C* is a tuning parameter; and *m* is the total number of term pairs in *P*. The formulation in equation 1 in Figure 2 finds a large-margin linear function that minimizes the number of pairs of training examples swapped with respect to their desired ranking order.

We chose SVM^{rank}[49], which implements rankSVM in an efficient way by using a cutting-plane algorithm and learns from large sparse data in linear time.

Figure 2. Objective function used in training ranking support vector machine.

$$\min_{w,\varepsilon_{i,j}\geq 0} w^T w + \frac{C}{m}\sum_{(i,j)\in P}\varepsilon_{i,j} \quad s.t.\ \forall(i,j)\in P: w^T(x_i - x_j)\geq 1-\varepsilon_{i,j} \qquad (1)$$

Baseline Features for Ranking

We implemented 9 features commonly used for KE [34,35,37,50,51].

Frequency-Based Features

The frequency-based features include term frequency, inverse document frequency, and TF-IDF. Term frequency is the number of occurrences of a candidate term in each individual EHR note. Inverse document frequency and TF-IDF are calculated in the

XSL•FO RenderX

standard way (see Multimedia Appendix 2). We used 6,237 clinical notes, which were selected by using the same 6 diagnoses used to select the 90 notes for the FOCUS corpus, to compute inverse document frequency.

Term Structure-Based Features

The term structure-based features include term length (TL) (ie, the total number of words contained in a term), the length of the longest word (by character) in a candidate term (maxWL), and a combined feature of TL and maxWL [51], as defined in equation 2 in Figure 3.

Since longer terms and words are less likely to be familiar to patients, these features may help distinguish between unfamiliar and common or familiar terms. Thus, these features may help rank as low EHR terms that are too common to be important (eg, *blood* and *pain*).

Figure 3. Equation for defining a combined feature of TL and maxWL. TL: term length (ie, length of a candidate term by word); maxWL: length of the longest word (by character) in a candidate term.

$$TL \times \max WL(t) = \sqrt{\log(1+TL(t)) \times \log(1+\max WL(t))} \qquad (2)$$

Position Feature

The position feature is the number of words preceding the first occurrence of a candidate term, normalized by the total number of words in the document. We used this feature because we found that the medical terms most specific to a patient often occur early in his/her EHR notes.

Lexical Feature

The lexical feature was found to be useful in domain-specific KE [35]. In our experiments, we used Porter’s stemmer to normalize terms. Since EHR data is noisy, we empirically include a stemmed term only if it occurs at least 3 times in the training data to eliminate misspelled words.

Part-of-Speech Feature

We used the POS tag of the head word of each candidate term, as generated by the clinical Text Analysis and Knowledge Extraction System (cTAKES) [52].

Additional Features for Ranking

Distributed Word Representation (Word Embedding)

Word embeddings are distributed vector representations of words learned from large unlabeled data. Words sharing similar semantics and context are expected to be close in their word vector space [53].

We include this feature because word embedding has emerged as a powerful technique for word representation. It has shown to improve several biomedical and clinical NLP tasks, such as biomedical named entity recognition [54,55], protein-protein interaction detection [56], biomedical event extraction [57,58], adverse drug event detection [59,60], ranking biomedical synonyms [61], and disambiguating clinical abbreviations [62,63].

We trained a neural language model to learn word embeddings. Specifically, we used Word2Vec software to create the skip-gram word embeddings [53,64]. We trained Word2Vec using a combined text corpus (over 3G words) of English Wikipedia, articles from PubMed Open Access, and 99,735 EHR notes from the Pittsburg corpus (Chapman W, University of Pittsburgh NLP Repository; using this data requires a license). We set the training parameters based on the study of Pyysalo et al [65]. We represented multi-word terms with the mean of individual word vectors. In this work, we used 200-dimension word vectors, with each dimension normalized to (0,1).

Unified Medical Language System Semantic Type

We mapped the candidate terms to Unified Medical Language System (UMLS) semantic types by using MetaMap, and included these semantic types as learning features.

Consumer Health Vocabulary Features

We derived 7 binary features from the consumer health vocabulary (CHV) [66]. The CHV is a collaborative resource and incorporates terms extracted from various consumer health sites, such as queries submitted to MedLinePlus and postings in health-focused online discussion forums [67-73]. The CHV contained 152,338 terms, most of which are consumer health terms [71-73]. Zeng et al [72] mapped these consumer health terms to the UMLS concepts by a semiautomatic approach. As a result of this work, the CHV encompasses lay terms as well as corresponding medical jargon.

In the FOCUS corpus, 89% of important terms are in the CHV, while a smaller percentage of nonimportant terms (76%) are in the CHV. This suggests that the presence of an EHR term in the CHV is indicative of the term’s importance from the perspective of patients (ie, health consumers). We therefore include a binary feature to denote whether a candidate term is in the CHV.

In addition, we derived 6 binary features from CHV familiarity scores. For extended usability, the CHV assigns familiarity scores to 57.89% (88,189/152,338) of its terms. CHV familiarity scores estimate the likelihood that a medical term can be understood by an average reader [74] and have values between 0 and 1, with 1 being most familiar and 0 being least familiar. CHV provides different types of familiarity scores [30]. Following Zeng-Treitler et al [30], we used the combined score and converted the continuous value into categorical features. Specifically, we divided the feature value range [0,1] into 5 equal-range bins, resulting in 5 binary features. The intuition behind these features is that medical terms with different levels of familiarity may be different in their importance to patients. For example, common terms (ie, terms that fall into the highest bin) such as *disease* and *physicians* are too general to be important. In addition, we included the sixth binary feature to indicate whether a candidate term has a CHV familiarity score.

XSL•FO
RenderX

Topic Features

Topic features are real-valued features in (0,1) to indicate the topic coherence between a candidate term and the EHR note containing this term. We compute topic features *P(t|e)* by equations 3 and 4 in Figure 4, where *P(t|e)* is the probability of a candidate term *t* conditioned on an EHR note *e*; *P(w|e)* is the probability of a word *w* conditioned on *e*; $P\ (w\mid topic_i)$ and $P\ (topic_i\mid e)$ are word-topic and topic-EHR note distributions estimated by the topic model; and *K* is the number of topics used in topic modeling.

We trained 3 latent Dirichlet allocation topic models with *K* set to 50, 100, and 200, respectively, after testing different *K*s on 6,237 clinical notes, which are the same as the notes used to compute IDF, using the MAchine Learning for LanguagE Toolkit (MALLET) [75] with default parameters to obtain 3 topic features.

Figure 4. Equations for defining topic feature.

$$P(w\mid e)=\sum_{i=1}^{K}P(w\mid topic_i)P(topic_i\mid e)\qquad(3)$$

$$P(t\mid e)=\sum_{w\in t}P(w\mid e)\qquad(4)$$

Training and Evaluation Settings

We created the training data from the FOCUS corpus as follows. We first applied MetaMap to the 90 notes in the FOCUS corpus. For each note, we took as positive examples those terms that were both identified by MetaMap and judged by physicians to be important to patients. We expanded the set of positive terms by using relaxed string match (details in the Evaluation Metrics subsection). The remaining terms identified by MetaMap were used as negative examples. This process resulted in a total of 690 positive and 21,809 negative terms from 90 notes.

Note that our 690 positive terms are less than the 793 terms annotated by physicians. This is because MetaMap missed some terms, many of which are multi-words with embedded UMLS concepts (eg, *autologous stem cell transplant* and *insulin-dependent diabetic*). Although we did not use these terms for training and for 10-fold cross-validation, we included them as positive terms for our final evaluation (as described in the Evaluation Metrics subsection).

We used the aforementioned training set for all the systems except 1 baseline system, adapted KEA++ (details in the Baseline Systems subsection), as it had its own procedure for extracting candidate terms and generating training data.

Previous work has shown that approximately 50-100 documents are sufficient to train supervised KE systems in the biomedical domain [45], suggesting that our 90 EHR notes, although a small size, may be sufficient. Our results empirically validated this hypothesis.

Baseline Systems

Adapted KEA++

The keyphrase extraction algorithm KEA [34] has been frequently used as a strong baseline in previous work [42,43,47]. KEA++ [45] is an extension of KEA with the added capacity for domain adaptation.

KEA++ is based on naïve Bayes and uses the following 4 features: TF-IDF, term position, term length in words, and a knowledge-based feature node degree. The last feature computes the number of semantic links in a knowledge base that connect a candidate phrase to other phrases in the document. In addition, it supports preselection and filtering of candidate terms by using controlled vocabularies, which we adapted to the clinical vocabularies.

Specifically, we included all the UMLS terms identified by MetaMap from the 90 FOCUS notes. We also included the complete list of medical terms from 3 comprehensive clinical vocabularies: MeSH, Systematized Nomenclature of Medicine (SNOMED), and the ninth revision of the International Classification of Diseases (ICD-9). To compute the node degree feature, we mapped terms in this controlled vocabulary to the UMLS concepts and incorporated concept relations (eg, *Is-a* and *Part-of*) from MeSH, SNOMED, and ICD-9.

Random Forest

RF [76] is an ensemble learning method that combines multiple decision trees for classification or regression. RF extends the idea of *bagging* [77] with a random selection of features [78-80] to improve robustness and generalizability. The RF classification method achieved the state-of-the-art performance—outperforming KEA and kernel SVMs—in extracting keyphrases from scientific literature [43].

We used the RF classification algorithm for our study. Assuming *t* is a candidate term from an EHR note *e*, the prediction of RF on (*t*, *e*), *f*(*t*,*e*), is calculated by equation 5 in Figure 5, where f_k(*t*,*e*) is the prediction on (*t*, *e*) (ie, the predicted possibility of *t* being an important medical term in *e*) by the *k*th decision tree among *B* decision trees built for RF (see more details below). According to equation 5 in Figure 5, *f*(*t*,*e*) represents the averaged predicted possibility of *t* being an important medical term in *e* and, therefore, can be used to rank candidate terms in *e*.

Each individual decision tree f_k is built as follows: assuming the training set contains *N* labeled examples (ie, *N* pairs of *t* and *e*, labeled as 1 if *t* is important in *e* and 0 if not) represented by *d* features, a single tree is built on *N* examples randomly sampled with replacement from this training set. When growing the tree, at each node the algorithm searches a randomly selected subset of the *d* features and selects 1 feature to create an if-then-else decision rule to branch the tree (ie, splitting the training examples at this node base on their feature values for the selected feature). Common criteria for selecting the feature that best splits a node include Gini impurity and information gain. When a node contains examples from the same class or its impurity is below a threshold, splitting stops and the node becomes a leaf node.

For a new example (*t*, *e*), RF assigns (*t*, *e*) to a leaf node of each individual decision tree by applying the decision rules learned from the training phase. The term $f_k(t,e)$ in equation 5 in Figure 5 is calculated as the fraction of positive training examples in the leaf node of the *k*th decision tree where (*t*, *e*) is assigned.

RF uses the same features as FOCUS. We used scikit-learn [81] to develop RF. We set the parameter *B* by minimizing the out-of-bag error during training and used default values for other parameters.

Figure 5. Prediction function of random forest.

$$f(t,e) = \frac{1}{B}\sum_{k=1}^{B} f_k(t,e) \qquad (5)$$

Evaluation Metrics

Precision, Recall, and F-score at Rank n

We report the averaged precision, recall, and *F*-score at ranks 5 and 10, abbreviated as P5, R5, and F5; and P10, R10, and F10, respectively. These metrics measure system performance for top ranks and are widely used to evaluate KE systems. We computed these metrics for the final evaluation (Step 3 in Figure 1) where we used all the gold-standard important terms as positive examples, including those that would never be included in the stage of candidate term extraction.

Area Under the Receiver Operating Characteristic Curve

Area under the receiver operating characteristic curve (AUC-ROC) is a metric widely used for evaluating ranking outputs. It computes the area under a receiver operating curve, which plots the true positive rate (y-coordinate) against the false positive rate (x-coordinate) at various threshold settings. To evaluate a system, we compute its AUC-ROC for each EHR note in the FOCUS corpus and report the averaged value. AUC-ROC measures the performance of the global ranking. Because both candidate term extraction and ranking affect the quality of global ranking, we report 2 AUC-ROC metrics: AUC-ROC$_{\text{ranking}}$ and AUC-ROC$_{\text{KE}}$. AUC-ROC$_{\text{ranking}}$ is computed on the candidate terms extracted by a system. Thereby, if a gold-standard important term is missed in candidate term extraction, it will not affect the system's AUC-ROC$_{\text{ranking}}$. Since this metric is informative about the ranking performance of a system, we used it to evaluate the cross-validation results on ranking candidate terms (Step 2.2 in Figure 1). AUC-ROC$_{\text{KE}}$ is computed by using all the gold-standard important terms as positive examples and measures the combined performance of candidate term extraction and ranking (Step 3 in Figure 1).

In the evaluation step, we use relaxed string match to determine true positives, as exact match is known to underestimate performance as perceived by human judges [50,82]. Specifically, we treat a term from the system output as a true positive if it either exactly matches or subsumes a gold-standard important term (eg, *non-Hodgkin lymphoma* subsumes *lymphoma*). We allow *subsume* but not *part-of* match in relaxed string match, as previous work found that the former aligned well with human judges but the latter did not [82]. For example, a part of an important term may be too general to be important (eg, *disease* in *Crohn's disease* and *iron* in *iron deficiency*).

Statistical Analysis

The paired samples *t* test was used for significance testing for the performance difference of 2 systems.

Results

Statistics of FOCUS Corpus

For each note, we treat the terms agreed by 2 physicians as the gold-standard important terms. In total, the physicians have identified 793 important medical terms from the 90 FOCUS notes (mean 9 [SD 5] terms per note). The Cohen's kappa coefficient for annotation agreement (microaverage) is .51. Table 1 summarizes the statistics of the FOCUS corpus.

The important terms identified by the physicians cover a wide range of topics, as represented by the UMLS semantic types. Table 2 shows term frequency and example terms for the 8 major topics.

XSL•FO
RenderX

Table 1. Statistics of the FOCUS[a] corpus.

Characteristics of the FOCUS corpus	*N* or mean (SD)
Number of notes, *N*	90
Number of words per EHR[b] note, mean (SD)	816 (133)
Number of candidate terms identified by MetaMap per EHR note, mean (SD)	250 (42)
Number of important medical terms identified by physicians per EHR note, mean (SD)	9 (5)

[a]FOCUS: Finding impOrtant medical Concepts most Useful to patientS.

[b]EHR: electronic health record.

Table 2. The 8 major topics in the FOCUS[a] corpus.

UMLS[b] semantic type	Number of important terms, *n*	Example terms
Disease or syndrome	295	autoimmune hemolytic anemia, gastroesophageal reflux, pancytopenia, Sjogren's syndrome, osteoporosis
Organic chemical	88	atenolol, vincristine, warfarin, Wellbutrin, Zocor
Finding	59	alopecia, hematuria, hypertension, NSTEMI (non-ST-elevation myocardial infarction), retinopathy
Neoplastic process	35	dermoid, large B cell lymphoma, pancreatic neoplasm, thyroid nodule
Therapeutic or preventive procedure	34	chemotherapy, dialysis, immunosuppression, kidney transplantation, pancreatectomy
Amino acid, peptide, or protein[c]	30	basal insulin, Rituxan, Neupogen, Synthroid, hemoglobin A1C, HPL (human placental lactogen)
Pathologic function	25	atrial fibrillation, autonomic dysfunction, BPH (benign prostatic hyperplasia), microscopic hematuria, systolic dysfunction
Diagnostic procedure	17	thyroid ultrasound, echocardiogram, endoscopy, biopsy, cardiac catheterization

[a]FOCUS: Finding impOrtant medical Concepts most Useful to patientS.

[b]UMLS: Unified Medical Language System.

[c]Electronic health record terms in this topic were split into 2 subtopics: medicine (denoted by their ingredients) and laboratory measure.

Most of the important terms annotated by physicians are specific to individual patients or notes. We used 2 criteria to select terms that may in general be important to patients: (1) the term occurs in more than 10% (9/90) of notes in the FOCUS corpus; and (2) the term was annotated as an important term for over 50% of the notes containing it. Only 4 terms were qualified and selected (the 2 bracketed numbers following the terms are the number of notes containing the term and the number of notes for which the term was annotated as important): *coronary artery disease* (20/14), *osteoarthritis* (19/10), *anemia* (13/7), and *prednisone* (10/6).

In addition, we made several observations from the FOCUS corpus. First, physicians typically excluded highly domain-specific terms that are very difficult for patients to understand. For example, the terms describing surgical procedures in detail or the anatomical parts of organs were excluded. Second, physicians often selected diseases and other information that are of immediate concern to patients, thus excluding other comorbidity diseases, for example.

Candidate Term Extraction

On average, adapted KEA++ extracts 342 candidate terms per note from the FOCUS corpus, which match 86% of the gold-standard physician annotated terms; FOCUS (the same for RF) extracts 250 candidates per note, which match 89% of the gold-standard terms.

Evaluation on FOCUS Corpus

Table 3 shows the evaluation results on the FOCUS corpus, where FOCUS achieves the best results and RF is the second best.

The performance difference between FOCUS and adapted KEA++ is statistically significant for all the metrics (P<.001). The difference between FOCUS and RF is also statistically significant for all the metrics (see *P* values in Table 3).

XSL•FO RenderX

Table 3. Performance of different natural language processing systems.

System	P5[a]	R5[b]	F5[c]	P10[d]	R10[e]	F10[f]	$\text{AUC-ROC}_{\text{ranking}}$[g]	$\text{AUC-ROC}_{\text{KE}}$[h]
Adapted KEA++[i]	0.333	0.211	0.239	0.281	0.362	0.292	0.890	0.780
RF[j]	0.409	0.267	0.299	0.339	0.416	0.346	0.891	0.821
FOCUS[k]	0.462	0.305	0.341	0.369	0.464	0.381	0.940	0.866
P (FOCUS vs RF)	.01	.01	.01	.045	.03	.02	<.001	<.001

[a] P5: precision at rank 5.

[b] R5: recall at rank 5.

[c] F5: *F*-score at rank 5.

[d] P10: precision at rank 10.

[e] R10: recall at rank 10.

[f] F10: *F*-score at rank 10.

[g] $\text{AUC-ROC}_{\text{ranking}}$: area under the receiver operating characteristic curve computed on the candidate terms extracted by a system.

[h] $\text{AUC-ROC}_{\text{KE}}$: area under the receiver operating characteristic curve (KE: keyphrase extraction) computed by using all the gold-standard important terms as positive examples.

[i] KEA++: extension of the keyphrase extraction algorithm KEA.

[j] RF: random forest.

[k] FOCUS: Finding impOrtant medical Concepts most Useful to patientS.

Textbox 2. Top-10 terms identified by different natural language processing systems for the full note containing the electronic health record excerpt in Textbox 1. True positives are italicized.

Adapted KEA++: *Crohn disease*, cirrhosis, *metformin*, recent, iron deficiency, *thrombocytosis*, Crohn, *diabetes mellitus*, anemia, omeprazole

RF (random forest): cirrhosis, iron deficiency anemia, iron deficiency, *thrombocytosis*, fenofibrate, alcohol, cheilosis, *Crohn disease, myeloproliferative neoplasms, metformin*

FOCUS (Finding impOrtant medical Concepts most Useful to patientS): *thrombocytosis, diabetes mellitus*, cirrhosis, *diabetes, metformin*, omeprazole, iron deficiency anemia, fenofibrate, *Crohn disease, budesonide*

Textbox 2 shows the top-10 terms identified by each of the 3 systems for the full note containing the EHR excerpt in Textbox 1 (where true positives are italicized). The $\text{AUC-ROC}_{\text{KE}}$ scores achieved by the 3 systems on the full note are 0.868 (FOCUS), 0.809 (adapted KEA++), and 0.857 (RF).

Effects of Additional Features

We tested the effects of the additional features on FOCUS and RF. The results (see Table 4) show that the additional features improve the performances of both FOCUS and RF substantially (FOCUS vs FOCUS-base and RF vs RF-base). The difference is statistically significant for all the metrics except R10 between RF and RF-base.

We further tested the effect of each additional feature by adding it on FOCUS-base. The results (see Table A3-1 in Multimedia Appendix 3) show that each additional feature improves the baseline features to a certain degree.

We then tested FOCUS's performance by using only additional features. The results (see Table A3-2 in Multimedia Appendix 3) show that word embedding is the best single feature, but still performs significantly worse than using all additional features for all the metrics (see row 5 in Table A3-2 in Multimedia Appendix 3 for *P* values). In addition, using only additional features performs significantly worse than using all features for all the metrics (*P*<.001).

XSL•FO
RenderX

Table 4. Performance of natural language processing systems with and without the additional features.

System	P5[a]	R5[b]	F5[c]	P10[d]	R10[e]	F10[f]	AUC-ROC$_{ranking}$[g]	AUC-ROC$_{KE}$[h]
FOCUS-base[i]	0.413	0.256	0.295	0.331	0.401	0.337	0.911	0.840
FOCUS[j]	0.462	0.305	0.341	0.369	0.464	0.381	0.940	0.866
P (FOCUS vs FOCUS-base)	.03	.02	.02	.003	<.001	.001	<.001	<.001
RF-base[k]	0.349	0.219	0.251	0.303	0.381	0.315	0.848	0.781
RF[l]	0.409	0.267	0.299	0.339	0.416	0.346	0.891	0.821
P (RF vs RF-base)	.003	.01	.01	.01	.10	.046	<.001	<.001

[a] P5: precision at rank 5.

[b] R5: recall at rank 5.

[c] F5: *F*-score at rank 5.

[d] P10: precision at rank 10.

[e] R10: recall at rank 10.

[f] F10: *F*-score at rank 10.

[g] AUC-ROC$_{ranking}$: area under the receiver operating characteristic curve computed on the candidate terms extracted by a system.

[h] AUC-ROC$_{KE}$: area under the receiver operating characteristic curve (KE: keyphrase extraction) computed by using all the gold-standard important terms as positive examples.

[i] FOCUS-base: Finding impOrtant medical Concepts most Useful to patientS; uses only the baseline features.

[j] FOCUS: Finding impOrtant medical Concepts most Useful to patientS; uses the baseline features plus the additional features.

[k] RF-base: random forest; uses only the baseline features.

[l] RF: random forest; uses the baseline features plus the additional features.

Discussion

Principal Findings

We have shown that physicians were able to identify important terms from EHR notes with moderate agreement (Cohen's kappa .51). This level of annotation agreement is acceptable for keyphrase annotation tasks [40,42,83]. We used the physicians' agreement to obtain high-quality data to develop and evaluate systems that automated this task.

Automated identification of EHR terms important to patients is challenging for several reasons. First, although frequency-based statistics such as term frequency and TF-IDF are widely used to estimate the importance of a term for a document, they are less effective for EHRs. For example, in our data, 56% of important medical terms occur only once in any individual EHR note. Second, we cannot infer the importance of a medical term solely based on its unfamiliarity level, as introduced in the Background and Significance subsection. Third, physicians' annotations cannot be represented by simple patterns. One reason is that most patients in our data have comorbidity and the important terms identified by physicians are usually related to only some of their diseases. In addition, the important terms are spread over a wide range of topics—details in the Statistics of FOCUS Corpus subsection—and thus cannot be inferred by manual categorical rules. Fourth, EHR notes contain abundant medical terms, among which only a small portion (4% in our case) were annotated as positive or important. Such imbalanced data pose extra challenges for supervised learning.

Despite the above challenges, our FOCUS system achieves a decent 0.866 AUC-ROC, suggesting that the learning-to-rank model with rich features is effective.

FOCUS Versus Adapted KEA++ and Random Forest

Our experiments show that FOCUS outperformed both adapted KEA++ and RF.

Using a more sophisticated MetaMap system, FOCUS is more effective than adapted KEA++ in candidate term extraction, as reported in the Candidate Term Extraction subsection. MetaMap is a state-of-the-art lexical tool that is well-configured—using morphological analysis and nonexact string match—to detect medical concepts and their corresponding medical terms from text, while adapted KEA++ uses a simpler approach (ie, dictionary look-up of stemmed *n*-grams from text).

We further compared FOCUS and adapted KEA++ on 28 FOCUS notes for which the 2 systems have the same recall on candidate extraction. FOCUS outperforms adapted KEA++ on this subset in all the evaluation measures, in particular, with significant improvements on AUC-ROC$_{ranking}$(0.936 vs 0.903, *P*=.03) and AUC-ROC$_{KE}$(0.875 vs 0.844, *P*=.03). This indicates that the rich features and the rankSVM algorithm contribute to FOCUS's performance gains.

Despite using the same MetaMap extractor and features, FOCUS still shows an advantage, outperforming RF in all the evaluation measures. The performance difference demonstrated that the ranking-based approach outperformed the state-of-the-art classification-based approach (RF) for this task. We attribute FOCUS's advantage over RF to the rankSVM algorithm used by FOCUS. Specifically, rankSVM sets its parameters by

XSL•FO
RenderX

minimizing the number of swapped pairs during its model training, which is equivalent to maximizing the rank quality as measured by Kendall's tau coefficient. In contrast, the RF algorithm is based on decision trees. The rules guiding the construction of decision trees (eg, information gain) are not directly optimizing rank quality.

We further analyzed the top-10 terms identified by the 3 systems. FOCUS, RF, and adapted KEA++ respectively ranked 433, 417, and 379 unique terms in their top-10 lists—since we have 90 notes, the maximum number of unique terms is 900. This result indicates that all 3 systems output diversified top-ranked terms, which are not constrained by a small set of terms, with FOCUS's output being the most diversified. We then identified terms frequently ranked as high (in the top 10) by each system using 2 criteria: (1) the term was identified as a candidate term for more than 10% (9/90) of the notes; and (2) the term was ranked in the top 10 over 60% of the time. The analysis results (see Table A4-1 in Multimedia Appendix 4) show that FOCUS and RF, RF and adapted KEA++, and FOCUS and adapted KEA++ share 6, 4, and 3 terms in their frequently ranked-as-high terms, respectively. Only 2 terms—*hypothyroidism* and *chemotherapy* —are frequently ranked as high by all 3 systems.

Effects of Additional Features

Our additional features, when applied jointly, improved both FOCUS and RF (see Table 4). As FOCUS and RF adopt different learning schemes—ranking versus classification—these results suggest that the beneficial effect of our additional features is generalizable to different learning methods.

Among the additional features, word embedding improves the AUC-ROC scores most—these scores measure the quality of the global ranking (see row 2 in Table A3-1 in Multimedia Appendix 3). This feature has been successfully applied to other biomedical and clinical NLP tasks. To the best of our knowledge, our work is the first to apply word embedding to ranking important terms in EHRs and show its usefulness.

The UMLS semantic type is the best in boosting performance at top ranks (rank=5 and rank=10, row 3 in Table A3-1 in Multimedia Appendix 3), suggesting its importance. One reason why it is useful is that medical terms with certain semantic types such as *medical device* and *anatomical structure* were almost never annotated by physicians as being important to patients. This feature, therefore, can help rank those terms lower to improve quality of top ranks.

Although the 3 topic features only improve the baseline features slightly, further analysis shows that they, when combined with other features, improve the performance. In particular, the FOCUS system using complete features significantly outperformed the one not using the topic features on AUC-ROC (*P*=.03 for both AUC-ROC$_{\text{ranking}}$ and AUC-ROC$_{\text{KE}}$).

The FOCUS systems that respectively use only all additional features and only word embedding achieved adequate results, especially on AUC-ROC scores (see Table A3-2 in Multimedia Appendix 3). However, they still performed worse than the system using all features, especially at top ranks.

Error Analysis and Future Work

We manually examined 17 notes, for which FOCUS has either zero recall at rank 5 or low AUC-ROC$_{\text{KE}}$(<0.800). We identified 3 error patterns.

First, we used relaxed string match for evaluation but did not allow *part-of* match, for the reason discussed in the Evaluation Metrics subsection. However, in some cases, this approach underestimates the performance. For example, FOCUS counted it as a mistake if MetaMap recognized *stem cell transplant* but not *autologous stem cell transplant*, the gold-standard term.

Second, FOCUS depends on MetaMap, which makes mistakes. It failed to identify certain abbreviations as medical terms (eg, *A1c* [a lab test for blood glucose], *BMD* [a lab test for bone mineral density], *CPPD* [calcium pyrophosphate deposition disease], and *TSH* [a lab test for thyroid stimulating hormone]). In future work, we may collect a list of common clinical abbreviations by mining a large EHR corpus and use this list to enhance medical term identification.

Third, the error is due to data sparsity. Although word embedding helps overcome data sparsity, FOCUS failed to rank as high some infrequent medical terms, such as *femoral popliteal bypass* and *pseudogout*. In future work, we will explore advanced approaches to deal with out-of-vocabulary words.

Limitations

Due to the common bottleneck of creating an expert-annotated resource, we only annotated 90 EHR notes for the reference standard and training data. Although this is not a large dataset, our system FOCUS shows an impressive performance of 0.940 AUC-ROC for 10-fold cross-validation on this data, suggesting that the data size may be sufficient.

Conclusions

We have presented a new clinical NLP task—identifying medical terms important to patients from EHRs. We developed FOCUS, a learning-based NLP system that is based on SVM learning-to-rank algorithm and rich learning features. The evaluation done on 90 physician-annotated EHR notes showed that FOCUS significantly outperformed other state-of-the-art NLP systems and that the additional features we developed were beneficial in boosting its performance.

Acknowledgments

This work was supported by the Investigator Initiated Research (1I01HX001457-01) from the Health Services Research and Development Program of the United States Department of Veterans Affairs. The content is solely the responsibility of the authors and does not represent the views of the United States Department of Veterans Affairs or the United States Government.

XSL•FO
RenderX

We thank the UMassMed annotation team, including Elaine Freund, Victoria Wang, Andrew Hsu, Barinder Hansra, and Sonali Harchandani, for creating the FOCUS corpus, and we thank Weisong Liu for technical support in collecting EHR notes. We also thank the anonymous reviewers for their constructive comments and suggestions.

Conflicts of Interest

None declared.

Authors' Contributions

HY and JC designed the study and led the effort to develop the annotation guideline. JC and JZ collected the data. JC designed and developed the FOCUS system, conducted the experiments, and drafted the manuscript. All authors have made significant contributions to data analysis and paper revision. HY is responsible for answering queries about the FOCUS corpus.

Multimedia Appendix 1

Guidelines for annotating medical terms important to patients in electronic health record notes.

[PDF File (Adobe PDF File), 454KB - medinform_v4i4e40_app1.pdf]

Multimedia Appendix 2

Formulas for calculating frequency-based features.

[PDF File (Adobe PDF File), 543KB - medinform_v4i4e40_app2.pdf]

Multimedia Appendix 3

Effects of additional features on FOCUS’s ranking performance. FOCUS: Finding impOrtant medical Concepts most Useful to patientS.

[PDF File (Adobe PDF File), 685KB - medinform_v4i4e40_app3.pdf]

Multimedia Appendix 4

Medical terms frequently ranked as high by different natural language processing systems.

[PDF File (Adobe PDF File), 450KB - medinform_v4i4e40_app4.pdf]

References

1. Vol Title XIII of Division A and Title IV of Division B of the American Recovery and Reinvestment Act of 2009. Washington, DC: Office of the National Coordinator for Health Information; 2009 Feb 18. Health Information Technology for Economic and Clinical Health Act (HITECH Act). URL: https://www.healthit.gov/sites/default/files/hitech_act_excerpt_from_arra_with_index.pdf [WebCite Cache ID 6m5sIHphk]
2. Steinbrook R. Health care and the American Recovery and Reinvestment Act. N Engl J Med 2009 Mar 12;360(11):1057-1060. [doi: 10.1056/NEJMp0900665] [Medline: 19224738]
3. Wright A, Feblowitz J, Samal L, McCoy AB, Sittig DF. The Medicare Electronic Health Record Incentive Program: Provider performance on core and menu measures. Health Serv Res 2014 Feb;49(1 Pt 2):325-346 [FREE Full text] [doi: 10.1111/1475-6773.12134] [Medline: 24359554]
4. Irizarry T, DeVito DA, Curran CR. Patient portals and patient engagement: A state of the science review. J Med Internet Res 2015;17(6):e148 [FREE Full text] [doi: 10.2196/jmir.4255] [Medline: 26104044]
5. Delbanco T, Walker J, Darer JD, Elmore JG, Feldman HJ, Leveille SG, et al. Open notes: doctors and patients signing on. Ann Intern Med 2010 Jul 20;153(2):121-125. [doi: 10.7326/0003-4819-153-2-201007200-00008] [Medline: 20643992]
6. HealthIT.gov. About the Blue Button movement. URL: https://www.healthit.gov/patients-families/about-blue-button-movement [accessed 2016-11-17] [WebCite Cache ID 6m5tZNgI8]
7. Delbanco T, Walker J, Bell SK, Darer JD, Elmore JG, Farag N, et al. Inviting patients to read their doctors' notes: A quasi-experimental study and a look ahead. Ann Intern Med 2012 Oct 2;157(7):461-470 [FREE Full text] [doi: 10.7326/0003-4819-157-7-201210020-00002] [Medline: 23027317]
8. Nazi KM, Hogan TP, McInnes DK, Woods SS, Graham G. Evaluating patient access to electronic health records: Results from a survey of veterans. Med Care 2013 Mar;51(3 Suppl 1):S52-S56. [doi: 10.1097/MLR.0b013e31827808db] [Medline: 23407012]
9. Woods SS, Schwartz E, Tuepker A, Press NA, Nazi KM, Turvey CL, et al. Patient experiences with full electronic access to health records and clinical notes through the My HealtheVet Personal Health Record Pilot: Qualitative study. J Med Internet Res 2013;15(3):e65 [FREE Full text] [doi: 10.2196/jmir.2356] [Medline: 23535584]

XSL•FO
RenderX

10. Zeng-Treitler Q, Kim H, Goryachev S, Keselman A, Slaughter L, Smith CA. Text characteristics of clinical reports and their implications for the readability of personal health records. Stud Health Technol Inform 2007;129(Pt 2):1117-1121. [Medline: 17911889]
11. Kandula S, Curtis D, Zeng-Treitler Q. A semantic and syntactic text simplification tool for health content. AMIA Annu Symp Proc 2010;2010:366-370 [FREE Full text] [Medline: 21347002]
12. Polepalli RB, Houston T, Brandt C, Fang H, Yu H. Improving patients' electronic health record comprehension with NoteAid. Stud Health Technol Inform 2013;192:714-718. [Medline: 23920650]
13. Sarzynski E, Hashmi H, Subramanian J, Fitzpatrick L, Polverento M, Simmons M, et al. Opportunities to improve clinical summaries for patients at hospital discharge. BMJ Qual Saf 2016 May 6. [doi: 10.1136/bmjqs-2015-005201] [Medline: 27154878]
14. Doak CC, Doak LG, Root JH. In: Morton PG, editor. Teaching Patients With Low Literacy Skills. 2nd edition. Philadelphia, PA: JB Lippincott Company; 1996.
15. Doak CC, Doak LG, Friedell GH, Meade CD. Improving comprehension for cancer patients with low literacy skills: Strategies for clinicians. CA Cancer J Clin 1998;48(3):151-162 [FREE Full text] [Medline: 9594918]
16. Walsh TM, Volsko TA. Readability assessment of Internet-based consumer health information. Respir Care 2008 Oct;53(10):1310-1315 [FREE Full text] [Medline: 18811992]
17. Eltorai AE, Han A, Truntzer J, Daniels AH. Readability of patient education materials on the American Orthopaedic Society for Sports Medicine website. Phys Sportsmed 2014 Nov;42(4):125-130. [doi: 10.3810/psm.2014.11.2099] [Medline: 25419896]
18. Morony S, Flynn M, McCaffery KJ, Jansen J, Webster AC. Readability of written materials for CKD patients: A systematic review. Am J Kidney Dis 2015 Jun;65(6):842-850. [doi: 10.1053/j.ajkd.2014.11.025] [Medline: 25661679]
19. Kutner M, Greenberg E, Jin Y, Paulsen C. The Health Literacy of America’s Adults: Results From the 2003 National Assessment of Adult Literacy. Washington, DC: US Department of Education, National Center for Education Statistics; 2006 Sep. URL: http://nces.ed.gov/pubs2006/2006483.pdf [accessed 2016-11-11] [WebCite Cache ID 6lwUr7mOK]
20. Pyper C, Amery J, Watson M, Crook C. Patients' experiences when accessing their online electronic patient records in primary care. Br J Gen Pract 2004 Jan;54(498):38-43 [FREE Full text] [Medline: 14965405]
21. Keselman A, Slaughter L, Smith CA, Kim H, Divita G, Browne A, et al. Towards consumer-friendly PHRs: Patients' experience with reviewing their health records. AMIA Annu Symp Proc 2007:399-403 [FREE Full text] [Medline: 18693866]
22. Chapman K, Abraham C, Jenkins V, Fallowfield L. Lay understanding of terms used in cancer consultations. Psychooncology 2003 Sep;12(6):557-566. [doi: 10.1002/pon.673] [Medline: 12923796]
23. Lerner EB, Jehle DV, Janicke DM, Moscati RM. Medical communication: Do our patients understand? Am J Emerg Med 2000 Nov;18(7):764-766. [doi: 10.1053/ajem.2000.18040] [Medline: 11103725]
24. Jones RB, McGhee SM, McGhee D. Patient online access to medical records in general practice. Health Bull (Edinb) 1992 Mar;50(2):143-150. [Medline: 1517087]
25. Baldry M, Cheal C, Fisher B, Gillett M, Huet V. Giving patients their own records in general practice: Experience of patients and staff. Br Med J (Clin Res Ed) 1986 Mar 1;292(6520):596-598 [FREE Full text] [Medline: 3081187]
26. Sarkar U, Karter AJ, Liu JY, Adler NE, Nguyen R, Lopez A, et al. The literacy divide: Health literacy and the use of an Internet-based patient portal in an integrated health system-Results from the diabetes study of northern California (DISTANCE). J Health Commun 2010;15 Suppl 2:183-196 [FREE Full text] [doi: 10.1080/10810730.2010.499988] [Medline: 20845203]
27. Zarcadoolas C, Vaughon WL, Czaja SJ, Levy J, Rockoff ML. Consumers' perceptions of patient-accessible electronic medical records. J Med Internet Res 2013;15(8):e168 [FREE Full text] [doi: 10.2196/jmir.2507] [Medline: 23978618]
28. Tieu L, Sarkar U, Schillinger D, Ralston JD, Ratanawongsa N, Pasick R, et al. Barriers and facilitators to online portal use among patients and caregivers in a safety net health care system: A qualitative study. J Med Internet Res 2015;17(12):e275 [FREE Full text] [doi: 10.2196/jmir.4847] [Medline: 26681155]
29. Aronson AR, Lang F. An overview of MetaMap: Historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
30. Zeng-Treitler Q, Goryachev S, Kim H, Keselman A, Rosendale D. Making texts in electronic health records comprehensible to consumers: A prototype translator. AMIA Annu Symp Proc 2007:846-850 [FREE Full text] [Medline: 18693956]
31. Abrahamsson E, Forni T, Skeppstedt M, Kvist M. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. In: Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PIT), 14th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics; 2014 Presented at: The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014); April 26-30, 2014; Gothenburg, Sweden p. 57-65 URL: http://www.aclweb.org/anthology/W14-1207
32. Zheng J, Yu H. Methods for linking EHR notes to education materials. AMIA Jt Summits Transl Sci Proc 2015;2015:209-215 [FREE Full text] [Medline: 26306273]
33. Hasan KS, Ng V. Automatic keyphrase extraction: A survey of the state of the art. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014). Stroudsburg, PA: Association for Computational

XSL•FO
RenderX

Linguistics; 2014 Presented at: The 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014); June 23-25, 2014; Baltimore, MD p. 1262-1273 URL: http://acl2014.org/acl2014/P14-1/pdf/P14-1119.pdf

34. Witten IH, Paynter GW, Frank E, Gutwin C, Nevill-Manning CG. KEA: Practical automatic keyphrase extraction. In: Proceedings of the Fourth ACM Conference on Digital Libraries.: ACM; 1999 Presented at: The Fourth ACM Conference on Digital Libraries; August 11-14, 1999; Berkeley, CA p. 254-255. [doi: 10.1145/313238.313437]
35. Frank E, Paynter GW, Witten IH, Gutwin C, Nevill-Manning CG. Domain-specific keyphrase extraction. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99). 1999 Presented at: The Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99); July 31-August 6, 1999; Stockholm, Sweden p. 668-673 URL: http://www.ijcai.org/Proceedings/99-2/Papers/002.pdf
36. Turney PD. Learning to Extract Keyphrases From Text. Ottawa, ON: National Research Council Canada, Institute for Information Technology; 1999 Feb 17. URL: http://extractor.com/ERB-1057.pdf [accessed 2016-11-10] [WebCite Cache ID 6lvC2cX9I]
37. Hulth A. Improved automatic keyword extraction given more linguistic knowledge. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics; 2003 Presented at: The 2003 Conference on Empirical Methods in Natural Language Processing; July 11-12, 2003; Sapporo, Japan p. 216-223 URL: http://www.aclweb.org/anthology/W03-1028
38. HaCohen-Kerner Y, Gross Z, Masa A. Automatic extraction and learning of keyphrases from scientific articles. In: Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'05). Berlin, Germany: Springer-Verlag; 2005 Presented at: The 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'05); February 13-19, 2005; Mexico City, Mexico p. 657-669. [doi: 10.1007/978-3-540-30586-6_74]
39. Yih W, Goodman J, Carvalho VR. Finding advertising keywords on Web pages. In: Proceedings of the 15th International Conference on World Wide Web (WWW '06). New York, NY: ACM; 2006 Presented at: The 15th International Conference on World Wide Web (WWW '06); May 23-26, 2006; Edinburgh, Scotland p. 213-222. [doi: 10.1145/1135777.1135813]
40. Medelyan O, Frank E, Witten IH. Human-competitive tagging using automatic keyphrase extraction. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009). Stroudsburg, PA: Association for Computational Linguistics; 2009 Presented at: 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009); August 6-7, 2009; Singapore p. 1318-1327. URL: http://www.cs.waikato.ac.nz/ml/publications/2009/maui_emnlp2009_1dataset.pdf
41. Lopez P, Romary L. HUMB: Automatic key term extraction from scientific articles in GROBID. In: Proceedings of the 5th International Workshop on Semantic Evaluation (ACL 2010). Stroudsburg, PA: Association for Computational Linguistics; 2010 Presented at: The 5th International Workshop on Semantic Evaluation (ACL 2010); July 15-16, 2010; Uppsala, Sweden p. 248-251 URL: http://www.aclweb.org/anthology/S10-1055
42. Jiang X, Hu Y, Li H. A ranking approach to keyphrase extraction. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM; 2009 Presented at: The 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval; July 19-23, 2009; Boston, MA p. 756-757 (see details in the Microsoft Research Technical Report MSR-TR-2009-96). URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.159.4470&rep=rep1&type=pdf
43. Krapivin M, Autayeu M, Marchese M, Blanzieri E, Segata N. Improving machine learning approaches for keyphrases extraction from scientific documents with natural language knowledge. In: Proceedings of the Joint JCDL/ICADL International Digital Libraries Conference (JCDL 2010). Berlin, Germany: Springer-Verlag; 2010 Presented at: The Joint JCDL/ICADL International Digital Libraries Conference (JCDL 2010); June 21-25, 2010; Gold Coast, Australia p. 102-111 URL: https://pdfs.semanticscholar.org/38f8/1d0a1eede4d7b7df169a92df22906c92a950.pdf
44. Li Q, Wu YF. Identifying important concepts from medical documents. J Biomed Inform 2006 Dec;39(6):668-679 [FREE Full text] [doi: 10.1016/j.jbi.2006.02.001] [Medline: 16545986]
45. Medelyan O, Witten IH. Domain-independent automatic keyphrase indexing with small training sets. J Am Soc Inf Sci Technol 2008 May;59(7):1026-1040. [doi: 10.1002/asi.20790]
46. Sarkar K. Automatic keyphrase extraction from medical documents. In: Proceedings of the 3rd International Conference on Pattern Recognition and Machine Intelligence (PReMI '09). 2009 Presented at: The 3rd International Conference on Pattern Recognition and Machine Intelligence (PReMI '09); December 16-20, 2009; New Delhi, India p. 273-278. [doi: 10.1007/978-3-642-11164-8_44]
47. Sarkar K. A hybrid approach to extract keyphrases from medical documents. Int J Comput Appl 2013 Feb 15;63(18):14-19. [doi: 10.5120/10565-5528]
48. Herbrich R, Graepel T, Obermayer K. Large margin rank boundaries for ordinal regression. In: Proceedings of Advances in Neural Information Processing Systems 1999 (NIPS 1999). 1999 Presented at: Advances in Neural Information Processing Systems 1999 (NIPS 1999); November 29-December 4, 1999; Denver, CO p. 115-132.
49. Joachims T. Training linear SVMs in linear time. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06). New York, NY: ACM; 2006 Presented at: The 12th ACM SIGKDD

XSL•FO
RenderX

International Conference on Knowledge Discovery and Data Mining (KDD'06); August 20-23, 2006; Philadelphia, PA p. 217-226. [doi: 10.1145/1150402.1150429]

50. Turney PD. Learning algorithm for keyphrase extraction. Inf Retr 2000;2(4):303-336. [doi: 10.1023/A:1009976227802]
51. Sarkar K, Nasipuri M, Ghose S. A new approach to keyphrase extraction using neural networks. Int J Comput Sci Issues 2010;7(2.3):16-25. [FREE Full text]
52. 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]
53. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of Advances in Neural Information Processing Systems 2013 (NIPS 2013). 2013 Presented at: Advances in Neural Information Processing Systems 2013 (NIPS 2013); December 5-10, 2013; Lake Tahoe, NV p. 3111-3119 URL: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
54. Tang B, Cao H, Wang X, Chen Q, Xu H. Evaluating word representation features in biomedical named entity recognition tasks. Biomed Res Int 2014;2014:240403 [FREE Full text] [doi: 10.1155/2014/240403] [Medline: 24729964]
55. Liu S, Tang B, Chen Q, Wang X. Effects of semantic features on machine learning-based drug name recognition systems: Word embeddings vs manually constructed dictionaries. Inf 2015 Dec 11;6(4):848-865. [doi: 10.3390/info6040848]
56. Jiang Z, Li S, Huang D. A general protein-protein interaction extraction architecture based on word representation and feature selection. Int J Data Min Bioinform 2016;14(3):276-291. [doi: 10.1504/IJDMB.2016.074878]
57. Li C, Song R, Liakata M, Vlachos A, Seneff S, Zhang X. Using word embedding for bio-event extraction. In: Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015). Stroudsburg, PA: Association for Computational Linguistics; 2015 Presented at: The 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015); July 30, 2015; Beijing, China p. 121-126 URL: http://www.aclweb.org/anthology/W15-3814
58. Nie Y, Rong W, Zhang Y, Ouyang Y, Xiong Z. Embedding assisted prediction architecture for event trigger identification. J Bioinform Comput Biol 2015 Jun;13(3):1541001. [doi: 10.1142/S0219720015410012] [Medline: 25669328]
59. Henriksson A, Kvist M, Dalianis H, Duneld M. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. J Biomed Inform 2015 Oct;57:333-349 [FREE Full text] [doi: 10.1016/j.jbi.2015.08.013] [Medline: 26291578]
60. Jagannatha AN, Yu H. Bidirectional RNN for medical event detection in electronic health records. In: Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics; 2016 Presented at: The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 12-17, 2016; San Diego, CA p. 473-482 URL: https://www.aclweb.org/anthology/N/N16/N16-1056.pdf
61. Jagannatha AN, Chen J, Yu H. Mining and ranking biomedical synonym candidates from Wikipedia. In: Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi). 2015 Presented at: The Sixth International Workshop on Health Text Mining and Information Analysis (Louhi); September 17, 2015; Lisbon, Portugal p. 142-151 URL: http://aclweb.org/anthology/W/W15/W15-2619.pdf
62. Wu Y, Xu J, Zhang Y, Xu H. Clinical abbreviation disambiguation using neural word embeddings. In: Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015). 2015 Presented at: The 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015); July 30, 2015; Beijing, China p. 171-176 URL: http://www.aclweb.org/anthology/W15-3822
63. Liu Y, Ge T, Mathews KS, Ji H, McGuinness DL. Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion. In: Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015). 2015 Presented at: The 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015); July 30, 2015; Beijing, China p. 92-97 URL: https://www.aclweb.org/anthology/W15-3810
64. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. ArXiv13013781 Cs; 2013 Jan 16. URL: https://www.researchgate.net/profile/Gs_Corrado/publication/234131319_Efficient_Estimation_of_Word_Representations_in_Vector_Space/links/5446726b0cf2f14fb80f3c7b.pdf?origin=publication_detail [WebCite Cache ID 6m6NhZqFz]
65. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional semantics resources for biomedical text processing. In: Proceedings of the 5th International Symposium on Languages in Biology and Medicine (LBM 2013). 2013 Presented at: The 5th International Symposium on Languages in Biology and Medicine (LBM 2013); December 12-13, 2013; Tokyo, Japan p. 39-43 URL: http://bio.nlplab.org/pdf/pyysalo13literature.pdf
66. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc 2006;13(1):24-29 [FREE Full text] [doi: 10.1197/jamia.M1761] [Medline: 16221948]
67. McCray AT, Loane RF, Browne AC, Bangalore AK. Terminology issues in user access to Web-based medical information. Proc AMIA Symp 1999:107-111 [FREE Full text] [Medline: 10566330]
68. Zeng Q, Kogan S, Ash N, Greenes RA. Patient and clinician vocabulary: How different are they? Stud Health Technol Inform 2001;84(Pt 1):399-403. [Medline: 11604772]

XSL•FO
RenderX

69. Patrick TB, Monga HK, Sievert ME, Houston HJ, Longo DR. Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes. J Med Internet Res 2001;3(3):e24 [FREE Full text] [doi: 10.2196/jmir.3.3.e24] [Medline: 11720966]
70. Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA. Characteristics of consumer terminology for health information retrieval. Methods Inf Med 2002;41(4):289-298. [Medline: 12425240]
71. Tse T, Soergel D. Exploring medical expressions used by consumers and the media: An emerging view of consumer health vocabularies. AMIA Annu Symp Proc 2003:674-678 [FREE Full text] [Medline: 14728258]
72. Zeng QT, Tse T, Crowell J, Divita G, Roth L, Browne AC. Identifying consumer-friendly display (CFD) names for health concepts. AMIA Annu Symp Proc 2005:859-863 [FREE Full text] [Medline: 16779162]
73. Keselman A, Smith CA, Divita G, Kim H, Browne AC, Leroy G, et al. Consumer health concepts that do not map to the UMLS: Where do they fit? J Am Med Inform Assoc 2008;15(4):496-505 [FREE Full text] [doi: 10.1197/jamia.M2599] [Medline: 18436906]
74. Zeng Q, Kim E, Crowell J, Tse T. A text corpora-based estimation of the familiarity of health terminology. In: Proceedings of the 6th International Symposium on Biological and Medical Data Analysis (ISBMDA 2005). 2005 Presented at: 6th International Symposium on Biological and Medical Data Analysis (ISBMDA 2005); November 10-11, 2005; Aveiro, Portugal p. 184-192 URL: https://lhncbc.nlm.nih.gov/files/archive/pub2005041.pdf
75. McCallum AK. MALLET: A Machine Learning for Language Toolkit. 2002. URL: http://mallet.cs.umass.edu [accessed 2016-07-04] [WebCite Cache ID 6il7RNCwf]
76. Breiman L. Random forests. Mach Learn 2001;45(1):5-32. [doi: 10.1023/A:1010933404324]
77. Breiman L. Bagging predictors. Mach Learn 1996;24(2):123-140. [doi: 10.1023/A:1018054314350]
78. Ho TK. Random decision forests. In: Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR'95). 1995 Presented at: The Third International Conference on Document Analysis and Recognition (ICDAR'95); August 14-15, 1995; Montreal, QC p. 278-282.
79. Amit Y, Geman D. Shape quantization and recognition with randomized trees. Neural Comput 1997 Oct;9(7):1545-1588. [doi: 10.1162/neco.1997.9.7.1545]
80. Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 1998;20(8):832-844. [doi: 10.1109/34.709601]
81. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825-2830. URL: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
82. Zesch T, Gurevych I. Approximate matching for evaluating keyphrase extraction. In: Proceedings of the 2009 International Conference on Recent Advances in Natural Language Processing (RANLP 2009). 2009 Presented at: The 2009 International Conference on Recent Advances in Natural Language Processing (RANLP 2009); September 14-16, 2009; Borovets, Bulgaria p. 484-489 URL: http://www.aclweb.org/anthology/R09-1086
83. Liu F, Pennell D, Liu F, Liu Y. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL. Stroudsburg, PA: Association for Computational Linguistics; 2009 Presented at: Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL; May 31-June 5, 2009; Boulder, CO p. 620-628 URL: http://www.anthology.aclweb.org/N/N09/N09-1070.pdf

Abbreviations

AUC-ROC: area under the receiver operating characteristic curve
BMD: bone mineral density
CBC: complete blood count
CHV: consumer health vocabulary
CPPD: calcium pyrophosphate deposition disease
cTAKES: clinical Text Analysis and Knowledge Extraction System
EHR: electronic health record
F5: *F*-score at rank 5
F10: *F*-score at rank 10
FOCUS: Finding impOrtant medical Concepts most Useful to patientS
ICD-9: ninth revision of the International Classification of Diseases
KE: keyphrase extraction
KEA: keyphrase extraction algorithm
KIP: keyphrase identification program
MALLET: MAchine Learning for LanguagE Toolkit
maxWL: length of the longest word (by character) in a candidate term
MeSH: Medical Subject Headings
NLP: natural language processing

XSL•FO
RenderX

P5: precision at rank 5
P10: precision at rank 10
POS: part of speech
R5: recall at rank 5
R10: recall at rank 10
rankSVM: ranking support vector machine
RF: random forest
SNOMED: Systematized Nomenclature of Medicine
SVM: support vector machine
TF-IDF: term frequency-inverse document frequency
TL: term length
TSH: thyroid stimulating hormone
UMLS: Unified Medical Language System

Edited by G Eysenbach; submitted 19.07.16; peer-reviewed by R Pivovarov, S Zhang; comments to author 18.08.16; revised version received 24.09.16; accepted 22.10.16; published 30.11.16.

Please cite as:
Chen J, Zheng J, Yu H
Finding Important Terms for Patients in Their Electronic Health Records: A Learning-to-Rank Approach Using Expert Annotations
JMIR Med Inform 2016;4(4):e40
URL: http://medinform.jmir.org/2016/4/e40/
doi:10.2196/medinform.6373
PMID:27903489

©Jinying Chen, Jiaping Zheng, Hong Yu. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 30.11.2016. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Web-based Real-Time Case Finding for the Population Health Management of Patients With Diabetes Mellitus: A Prospective Validation of the Natural Language Processing–Based Algorithm With Statewide Electronic Medical Records

Le Zheng[1,2*], BS; Yue Wang[2,3*], PhD; Shiying Hao[2*], PhD; Andrew Y Shin[2*], MD; Bo Jin[4], MS; Anh D Ngo[4], MD, DrPH; Medina S Jackson-Browne[4], PhD; Daniel J Feller[4], BA; Tianyun Fu[4], BS; Karena Zhang[2], NA; Xin Zhou[5], MD; Chunqing Zhu[4], MS; Dorothy Dai[4], BS; Yunxian Yu[6], MD PhD; Gang Zheng[3], PhD; Yu-Ming Li[5], MD; Doff B McElhinney[2], MD; Devore S Culver[7], MM; Shaun T Alfreds[7], MBA; Frank Stearns[4], MHA; Karl G Sylvester[2], MD; Eric Widen[4], MHA; Xuefeng Bruce Ling[2,6], PhD

[1]Tsinghua University, Beijing, China

[2]Stanford University, Stanford, CA, United States

[3]Zhejiang University, Hangzhou, China

[4]HBI Solutions Inc, Palo Alto, CA, United States

[5]Tianjin Key Laboratory of Cardiovascular Remodeling and Target Organ Injury, Pingjin Hospital Heart Center, Tianjin, China

[6]School of Medicine, Zhejiang University, Hangzhou, China

[7]HealthInfoNet, Portland, ME, United States

[*]these authors contributed equally

Corresponding Author:
Xuefeng Bruce Ling, PhD
Stanford University
S370 Grant Bldg
Stanford, CA,
United States
Phone: 1 650 427 9198
Fax: 1 650 723 1154
Email: bxling@stanford.edu

Abstract

Background: Diabetes case finding based on structured medical records does not fully identify diabetic patients whose medical histories related to diabetes are available in the form of free text. Manual chart reviews have been used but involve high labor costs and long latency.

Objective: This study developed and tested a Web-based diabetes case finding algorithm using both structured and unstructured electronic medical records (EMRs).

Methods: This study was based on the health information exchange (HIE) EMR database that covers almost all health facilities in the state of Maine, United States. Using narrative clinical notes, a Web-based natural language processing (NLP) case finding algorithm was retrospectively (July 1, 2012, to June 30, 2013) developed with a random subset of HIE-associated facilities, which was then blind tested with the remaining facilities. The NLP-based algorithm was subsequently integrated into the HIE database and validated prospectively (July 1, 2013, to June 30, 2014).

Results: Of the 935,891 patients in the prospective cohort, 64,168 diabetes cases were identified using diagnosis codes alone. Our NLP-based case finding algorithm prospectively found an additional 5756 uncodified cases (5756/64,168, 8.97% increase) with a positive predictive value of .90. Of the 21,720 diabetic patients identified by both methods, 6616 patients (6616/21,720, 30.46%) were identified by the NLP-based algorithm before a diabetes diagnosis was noted in the structured EMR (mean time difference = 48 days).

Conclusions: The online NLP algorithm was effective in identifying uncodified diabetes cases in real time, leading to a significant improvement in diabetes case finding. The successful integration of the NLP-based case finding algorithm into the Maine HIE

XSL•FO
RenderX

database indicates a strong potential for application of this novel method to achieve a more complete ascertainment of diagnoses of diabetes mellitus.

(JMIR Med Inform 2016;4(4):e37) doi:10.2196/medinform.6328

KEYWORDS

electronic medical record; natural language processing; diabetes mellitus; data mining

Introduction

Diabetes mellitus (DM) is a leading cause of mortality and morbidity and accounts for significant burden of disease worldwide [1,2]. In the United States, 9.3% of the population or 29.1 million people were reported to have diabetes in 2013, plus an estimate of 8.1 million people with undiagnosed diabetes [3,4]. Diabetes is a metabolic disorder caused by a high concentration of glucose in the blood stream. If untreated, diabetic patients will eventually develop a range of complications. Diabetes complications can be prevented through timely application of several measures such as lifestyle modification and control of blood glucose and blood pressure for diabetic patients [3,5-8].

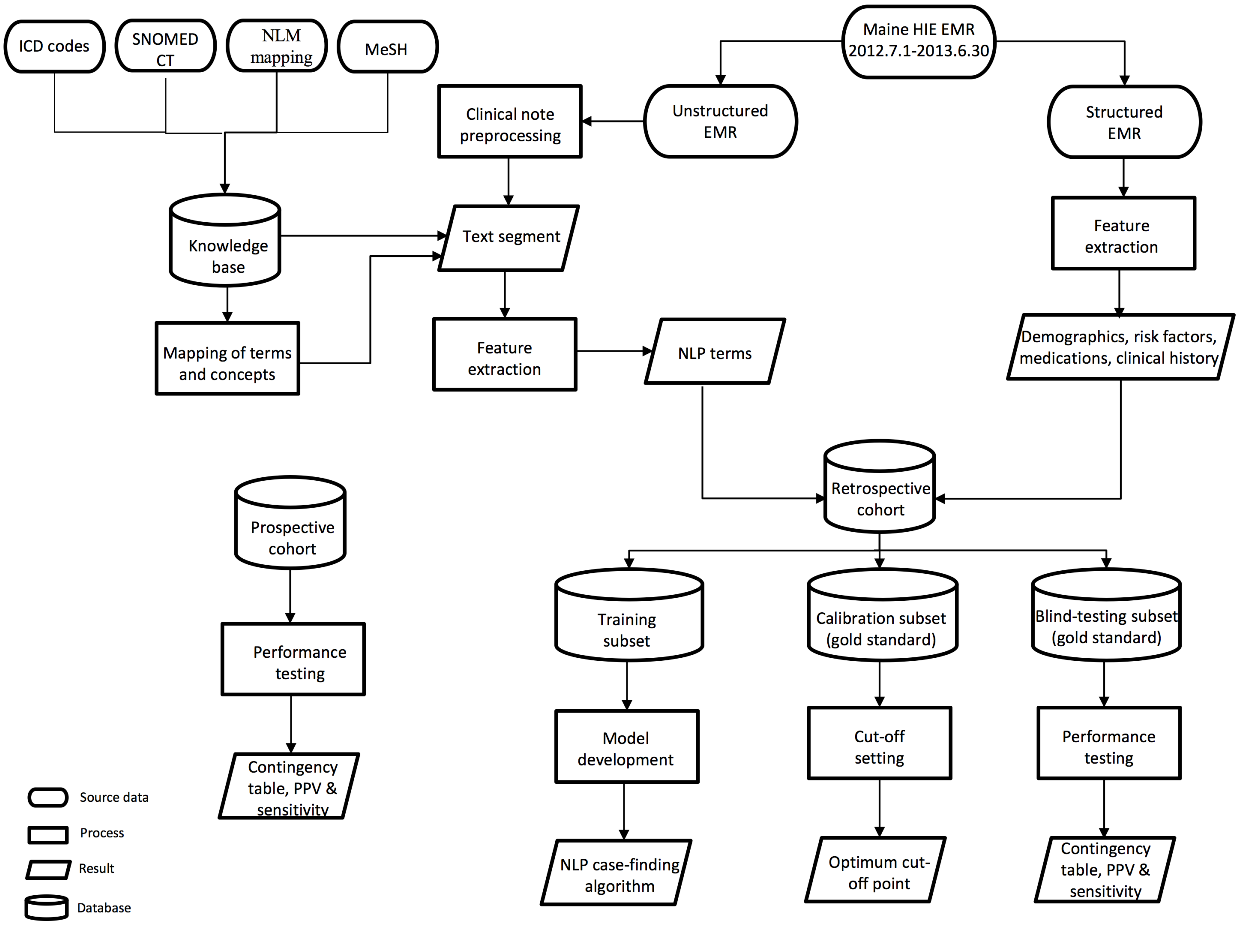
The identification of persons with diagnosed DM in electronic medical records (EMRs) is essential to quality improvement initiatives, clinical decision support systems, and regional disease prevalence estimates used by public health departments. Although DM diagnoses have typically been captured by International Classification of Diseases (ICD) codes and stored in EMRs, previous studies found that diagnostic codes alone do not adequately represent DM diagnoses across a population, resulting in underestimates of disease prevalence and challenging the development of electronic approaches to clinical management [9,10]. The prevalence of DM in 2014 in Maine was 7.8%, whereas the codified prevalence is 6.8% in our database. It indicates a gap caused by uncodified DM in the structured EMRs of patients. Diabetic patients who have received little or no diabetes care are unlikely to be associated with a diabetes-specific diagnosis code for billing, as are patients who transfer their care between multiple unaffiliated health care systems but receive no DM care for some time. To overcome this shortcoming, manual chart reviews of unstructured clinical notes have been used to identify uncodified DM cases. However, this method involves high labor costs and long latency, which has limited use for large scale datasets [11-13].

One possible solution to the problem and a fully automated alternative and acceptable means of delivering cost-effective case finding is the use of natural language processing (NLP), a Web-based technique. NLP has increasingly been used to enhance case finding for some high-impact chronic diseases such as heart failure and cancer through analyzing narrative text in EMRs [14-16]. The advantage of the automated NLP-based case finding algorithm is that it allows for the rapid real-time identification of uncodified diagnoses from large datasets. It also allows for the rapid preprocessing of unstructured clinical notes for different diseases and clinical conditions before a diagnosis is selected [14,16]. However, the existing NLP applications are mainly based on a small sample of patients with a limited number of clinical notes. Currently, the application of NLP in public health and medicine faces the following challenges [17-21]: (1) a lack of a comprehensive knowledge base to generate the accumulated domain knowledge from the targeted patient population; (2) a lack of a comprehensive data model to encapsulate the unstructured clinical notes of various formats across different health care facilities; (3) and a lack of a robust and scalable analytics pipeline to process a large number of EMR notes across statewide health care facilities.

The aim of this study was therefore to develop and integrate an online real-time NLP-based DM case finding algorithm into the health information exchange (HIE) care flow in the state of Maine, United States (Figure 1). We hypothesized that the algorithm we developed could find additional patients with DM who were not identified by codified diagnoses in structured EMRs. This algorithm was built on a knowledge base that incorporates taxonomies and controlled vocabularies encoding domain knowledge, as well as the task-oriented characteristics of clinical notes. It also used both structured and unstructured information and data available in EMRs, which were treated as variables for statistical learning in identification of uncodified DM diagnoses.

XSL•FO
RenderX

Figure 1. A schematic presentation of the natural language processing (NLP)–based algorithm integrated into the statewide diabetes mellitus case finding and surveillance. The clinical note was preprocessed and identified to generate the decision. The knowledge bases, statistical model, and the gold standard datasets form the basis of the NLP engine. ICD: International Classification of Diseases; NLM: US National Library of Medicine; MeSH: Medical Subject Headings; EMR: electronic medical record; HIE: health information exchange; PPV: positive predictive value. SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms.


ICD codes
SNOMED CT
NLM mapping
MeSH
Maine HIE EMR 2012.7.1-2013.6.30
Clinical note preprocessing
Unstructured EMR
Structured EMR
Knowledge base
Text segment
Feature extraction
Mapping of terms and concepts
Feature extraction
NLP terms
Demographics, risk factors, medications, clinical history
Retrospective cohort
Prospective cohort
Training subset
Calibration subset (gold standard)
Blind-testing subset (gold standard)
Performance testing
Model development
Cut-off setting
Performance testing
Contingency table, PPV & sensitivity
NLP case-finding algorithm
Optimum cut-off point
Contingency table, PPV & sensitivity
Source data
Process
Result
Database

Methods

Ethics Statements

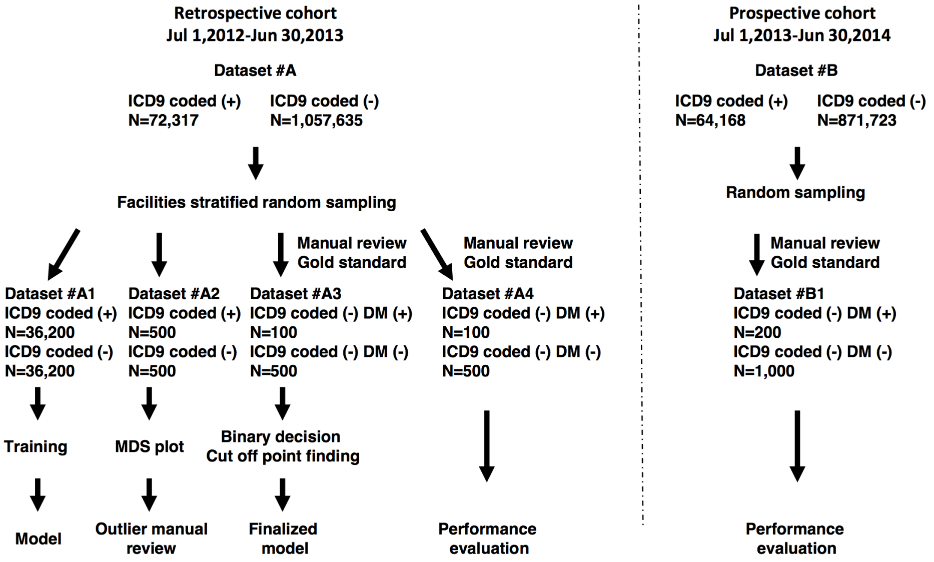
Protected personal health information was removed for the purpose of this research. Because this study analyzed deidentified data, it was exempted from ethics review by the Stanford University Institutional Review Board (October 16, 2014).

Data Sources

Data for this study were extracted from the HIE dataset administered by HealthInfoNet—an independent nonprofit organization. The dataset contains records of nearly 95% of the population in the state of Maine. There are 35 HIE-associated hospitals, 34 federally qualified health centers, and more than 400 ambulatory practices [22,23]. To identify the DM cohort, clinical notes of all categories in the Maine HIE EMR database were analyzed. Clinical notes are also known as progress notes, which are the part of a medical record where health care professionals document the details of a patient's clinical status or achievements during the course of inpatient care or outpatient care. Clinical notes in our study are encounter based. These notes were divided into 2 subcohorts. The retrospective cohort contained 1,385,280 notes representing 1,129,952 patients covering the period from July 1, 2012, to June 30, 2013, and the prospective cohort comprised 982,211 clinical notes representing 935,891 patients recorded from July 1, 2013, to June 30, 2014 (Figure 2). Clinical notes were derived from more than 100 different types of clinical reports, including history or physical reports, discharge summaries, and emergency reports.

XSL•FO
RenderX

Figure 2. Cohort construction of the study. ICD9: International Classification of Diseases, Ninth Revision; DM: diabetes mellitus; MDS: multidimensional scaling.


Retrospective cohort
Jul 1,2012-Jun 30,2013
Dataset #A
ICD9 coded (+) N=72,317
ICD9 coded (-) N=1,057,635
Facilities stratified random sampling
Manual review Gold standard
Manual review Gold standard
Dataset #A1 ICD9 coded (+) N=36,200 ICD9 coded (-) N=36,200
Dataset #A2 ICD9 coded (+) N=500 ICD9 coded (-) N=500
Dataset #A3 ICD9 coded (-) DM (+) N=100 ICD9 coded (-) DM (-) N=500
Dataset #A4 ICD9 coded (-) DM (+) N=100 ICD9 coded (-) DM (-) N=500
Training
MDS plot
Binary decision Cut off point finding
Model
Outlier manual review
Finalized model
Performance evaluation
Prospective cohort
Jul 1,2013-Jun 30,2014
Dataset #B
ICD9 coded (+) N=64,168
ICD9 coded (-) N=871,723
Random sampling
Manual review Gold standard
Dataset #B1 ICD9 coded (-) DM (+) N=200 ICD9 coded (-) DM (-) N=1,000
Performance evaluation

Algorithm Overview

The patients with DM were defined as those who had DM noted as either primary or secondary diagnosis (International Classification of Diseases, Ninth Revision, Clinical Modification, ICD-9-CM, codes: 249, 249.x, 249.xx, 250, 250.x, and 250.xx) in their medical records [24]. The case finding algorithm consisted of 3 sequential steps based on both structured and unstructured EMR information (Figure 1). The first step involved a preprocessing of unstructured clinical notes to remove information indicating the patient did not have DM, such as family history of DM and negation (ie, the patient denied DM). This step removed the misleading information to avoid false-positive errors, thus improving the performance of subsequent steps. The second step entailed a feature extraction that mapped DM risk factors recognized in previous studies [25-29], medications extracted from Unified Medical Language System, and NLP terms into the structured metadata. In the third step, a decision tree–based model based on the retrospective cohort was developed to determine whether a patient had DM. The development procedures are detailed in later sections. To support the whole algorithm pipeline, the NLP engine was created, including knowledge base, statistical models, and gold standard datasets as functional modules. Their construction and utilization are described below.

Knowledge Base

The knowledge base consisted of 3 cores: (1) DM-related clinical terms as the controlled vocabularies; (2) antidiabetic medications; and (3) extracted rules in the clinical notes.

Clinical terms in our NLP knowledge base were derived from the following sources: (1) the description and synonyms of ICD-9-CM codes under 249, 249.x, 249.xx, 250, 250.x, and 250.xx; (2) the comprehensive clinical terminologies within SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) [30]; (3) a mapping of ICD-9-CM with SNOMED CT proposed by the US National Library of Medicine (NLM) [31], based on the concepts and synonyms mapped to ICD codes 249, 249.x, 249.xx, 250, 250.x, and 250.xx; (4) the headings returned by the query of “diabetes” using NLM for article indexing [32] in a controlled vocabulary thesaurus, namely, Medical Subject Headings (MeSH). These clinical terms in the knowledge base were further tokenized, combined, and filtered to derive our controlled vocabulary of single and dual tokens. If those controlled vocabularies contained stop words, for example, “the,” “a,” “of,” provided by the text mining (tm) package (R Development Core Team) [33], they were removed. In total, 742 final NLP terms were identified (Multimedia Appendix 1); of these, 72 were found to be significantly associated with DM diagnosis (Mann-Whitney test *P* value <.05) in the retrospective cohort. Here, the patients who were assigned any of the ICD-9-CM codes 249, 249.x, 249.xx, 250, 250.x, 250.xx during the encounter were defined as having a diagnosis of DM.

Antidiabetic medications were identified from the Unified Medical Language System database. Out of 36 medications analyzed, 22 were found to be significantly associated with DM diagnosis (Mann-Whitney test *P* value <.05) in the retrospective cohort.

Because information on DM risk factors (eg, body mass index or BMI, high blood pressure, obesity, smoking history, and alcohol use disorders) might be presented in multiple unstructured formats in EMRs, we developed a series of regular expressions and rules to unify unstructured information and subsequently standardize feature categories. For example, BMI could be available from clinical notes, but in many instances only height and weight were provided. The BMI was then divided into 4 categories: underweight, normal, overweight, and obesity, according to the World Health Organization classification [34]. Additionally, to make the knowledge base more compatible with the expression of clinical notes, it was updated iteratively along with development of the retrospective model.

XSL•FO
RenderX

Preprocessing and Feature Extraction

Intuitively, DM-related words in the notes can be used to classify a DM case. However, this simpleminded note-processing method ignores negative expressions, for example, “The patient denied DM” in the note. Obviously, such negation will mislead the algorithm to wrongly classify the patient as a DM case. To avoid this kind of error, negation should be handled first before being fed into the pipeline. Preprocessing to remove family DM history is done because of similar considerations: the note with sentence “his mother had diabetes mellitus” does not classify the corresponding patient, “he,” as a diabetic patient. To ensure NLP specificity, segments associated with negation and family history of DM as described above were removed during preprocessing according to the entries in the knowledge base. The vocabulary of negation was populated using the lexicon proposed by NegEX [35]. The family-related words [36] were used to initiate the vocabulary of family history.

To break narrative text in clinical notes into smaller pieces, we applied the text semantics. A note was collapsed into paragraphs, sentences, and lines as basic units with nonoverlapping contents. Criteria to define a basic unit were developed on statistics of the text lengths and newline characters. If a paragraph (or a sentence, a line) satisfied criteria of a basic unit, it was regarded as one segment without further decomposition. The parts of speech were annotated and referred for sentence boundary detection against the confusion between periods and decimal points using openNLP (R Development Core Team) [33]. When a segment contained a word or a phrase in the vocabularies associated with negation and family history, this segment was removed from the note.

To map the unstructured text into structured metadata, the knowledge base was applied to the standardized clinical notes after preprocessing. When matching the text with the NLP terms and medications in the knowledge base was successful, the structured data of the notes were coded as “1,” otherwise as “0.” Then DM risk factors were extracted to further enrich the clinical notes metadata using the rules and regular expressions stored in the knowledge base.

Workflow of Gold Standard Dataset

Gold standard datasets were created for model development and validation purposes (Figure 2). The datasets contained a subset of patients with or without DM. The patient DM status was determined by manual chart reviews of clinical notes conducted by 2 physician-curators. If a patient had any notes showing DM diagnosis, he or she was coded as having DM. The 2 physicians reviewed each note individually and assessed whether the note showed the presence of DM. After individual review, the 2 assessments for each note were compared. Any disagreement was discussed by the 2 physicians and an agreement was reached [37]. When there was a disagreement on diagnosis that could not be resolved by discussion between the 2 curators, the patient was excluded. The datasets created through this process were used as the gold standard to define the cutoff point, run the blind testing, or to validate our NLP-based case finding algorithm. The cohort construction of the gold standard datasets is shown in Figure 2.

Model Development

A model was developed on the retrospective cohort (Figure 2). The clinic’s facilities where clinical notes were derived were randomly allocated to 1 of the 2 subsets: one for training and for finding the cutoff point (n=17 facilities) and the other for blind testing (n=18 facilities). Within the subsets for training and finding the cutoff point, all available notes (n=44,368) with codified DM diagnoses, and an equal number of uncodified notes (n=44,368), were selected to construct a training subcohort for model development. In the remaining uncodified subset, a gold standard dataset was constructed by randomly selecting 100 positive (DM) patients and 500 negative (non-DM) patients as the subcohort for finding the cutoff point. A further random sample of 100 positive and 500 negative patients identified from uncodified notes in the blind testing subset were selected to construct the blind testing subcohort.

By feeding the training subcohort to the preprocessing and feature extraction, each note had a feature vector denoted as *f*. The identification of DM was stated as maximum a posteriori probability (MAP) estimation in Figure 3 (a), where *DM* was a binary random variable indicating whether the sample had a DM diagnosis (*DM*=1). To take diagnosis codes into consideration, a binary variable *ICD* was introduced to indicate whether a note was codified (*ICD*=1). By inserting *ICD* into the posterior and then applying the Bayesian rule, we had the decomposition in Figure 3 (b).

Because the assignment of diagnosis code was independent of the extracted feature, the model was simplified to the equation in Figure 3 (c).

The first term on the right side determined the probability of DM for a codified note, while the second term on the right side for an uncodified note. As coding information was known, we had 2 branches to obtain the posterior a shown in Figure 3 (d).

The great majority of uncodified notes did not include a DM diagnosis, while most DM codified notes were ICD-9-CM DM diagnoses. This led us to develop the following class labeling method:

1. If a note is codified, this note should have a diagnosis of DM (Figure 3 (e));

2. If a note is not codified, a model should be built to estimate the probability (Figure 3 (f)).

As a result, the inference of DM diagnosis for a codified note was only dependent on the ICD code noted in the structured data, whereas for uncodified notes we trained a random forest model [33,38] to obtain T(*f*) (Figure 3 (g)), where t_n was the *n* th decision tree in the random forest.

At the perspective of hierarchical tree, the model could be considered as a combination of a predetermined tree-based model and a random forest-based model. The predetermined tree was developed based on the ICD-9-CM diagnosis codes associated with DM, which represented human prior knowledge. The random forest-based model was developed by extracting information from clinical notes, which represented machine learning knowledge.

XSL•FO
RenderX

The model was first trained with codified notes, the DM-positive sample, and uncodified notes, the DM-negative sample. The false positives in the training sample were uncodified notes either with or without a DM diagnosis. The former was regarded as the positive sample in the next round of training. By applying the 2 steps iteratively, the model as well as the knowledge base associated with the expression of family history and negation was fine-tuned. All false-positive cases were reviewed manually to understand how these occurred.

This codified-note–driven iterative training scheme was based on the hypothesis that the notes' features should be similar between codified notes and uncodified notes where a DM diagnosis was found. To test this hypothesis and validate the method, multidimensional scaling (MDS) plots were constructed with 1000 samples randomly selected from the training subcohort to illustrate the distribution of notes.

Figure 3. Equations describing the modeling process of the natural language processing (NLP)–based algorithm.

(a) $$\overline{DM} = \underset{DM}{\operatorname{argmax}}\, \mathrm{P}(DM|f)$$

(b) $$\mathrm{P}(DM|f) = \mathrm{P}(DM|ICD = 1, f)\mathrm{P}(ICD = 1|f) + \mathrm{P}(DM|ICD = 0, f)\mathrm{P}(ICD = 0|f)$$

(c) $$\mathrm{P}(DM|f) = \mathrm{P}(DM|ICD = 1, f)\mathrm{P}(ICD = 1) + \mathrm{P}(DM|ICD = 0, f)\mathrm{P}(ICD = 0)$$

(d) $$\mathrm{P}(DM|f) = \begin{cases} \mathrm{P}(DM|ICD = 1, f) & \text{codified} \\ \mathrm{P}(DM|ICD = 0, f) & \text{uncodified} \end{cases}$$

(e) $$\mathrm{P}(DM = 1|f) = \mathrm{P}(DM = 1|ICD = 1, f) = 1$$

(f) $$\mathrm{P}(DM = 1|f) = \mathrm{P}(DM = 1|ICD = 0, f) = \mathrm{T}(f)$$

(g) $$\mathrm{T}(f) = \frac{1}{N}\sum_{n} t_n(f)$$

Patient Classification Cutoff Point Determination

As the algorithm was developed to find out uncodified DM cases, the proportion of true positives among the identified samples, positive predictive value (PPV), was the most important indicator of performance. With a PPV of ≥90%, the proportion of false-positive cases is less than 10%. On the other hand, given that the method was to identify uncodified cases in addition to the codified cases, maintaining a high level of PPV at the expense of sensitivity is acceptable. The way we located the optimal cutoff by considering the trade-off between PPV and sensitivity was also presented in a previous NLP study [39]. Given that our algorithm assigned a classification probability to each subject, we aimed to find an optimal cutoff point to achieve the maximum classification sensitivity with a predefined PPV of 90%. To achieve a 90% PPV, the classification specificity can be calculated through a linear formula, thus forming a straight line overlaid on the receiver operating characteristic (ROC) curve. The combination of sensitivity and specificity in the region above the line allowed for a performance with >90% PPV. Thus, the cutoff point was set at the first intersection between the line and the ROC curve.

At the final stage of the retrospective model development, the case finding algorithm was blind tested on patients from health care facilities that were not included in the training subset.

Prospective Case Finding and Validation

Our NLP-based DM case finding algorithm was then deployed online through integration into the HIE real-time population exploration dashboard system. The clinical notes (N=982,211) covering the period from July 1, 2013, to June 30, 2014, were aggregated for prospective validation of the algorithm. An independent gold standard dataset was constructed based on chart reviews of clinical notes of 200 patients with DM and 1000 patients without DM randomly selected from the prospective cohort (Figure 2). The prospective classification performance on the gold standard dataset was evaluated using the following parameters: PPV, sensitivity, specificity, negative predictive value (NPV), and the area under the ROC curve. A total of 200 samples were further randomly selected from the uncodified DM cases identified by the algorithm to evaluate the case finding accuracy on the entire prospective cohort. On the basis of the longitudinal records of both clinical notes and diagnosis codes for each patient in the HIE EMR database, a temporal comparison of the 2 sources was analyzed.

Results

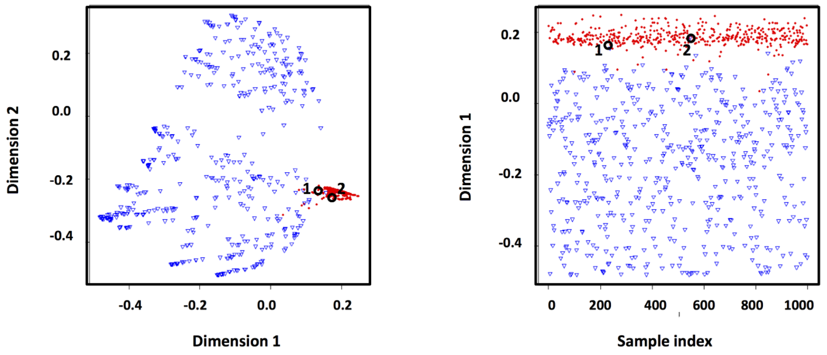
Case Finding Algorithm Performance

An MDS plot was constructed to visualize the classification performance. As shown in Figure 4, out of 500 uncodified notes, 2 were classified as DM diagnosis. A closer examination revealed that these "false-positive" cases had notes with genuine diagnosis of DM. This MDS plot indicated that (1) our model effectively differentiated the notes from those patients with DM diagnosis and those without DM diagnosis and (2) our NLP-based analysis of clinical notes can identify uncodified notes with diagnosis of DM.

Figure 4 shows that more than 99% of the uncodified notes were linked to patients without DM diagnosis and more than 99% of the codified notes were linked to patients with DM diagnosis. There were only 1% mislabeled samples in the training dataset, which did not alter the model performance [40].

XSL•FO
RenderX

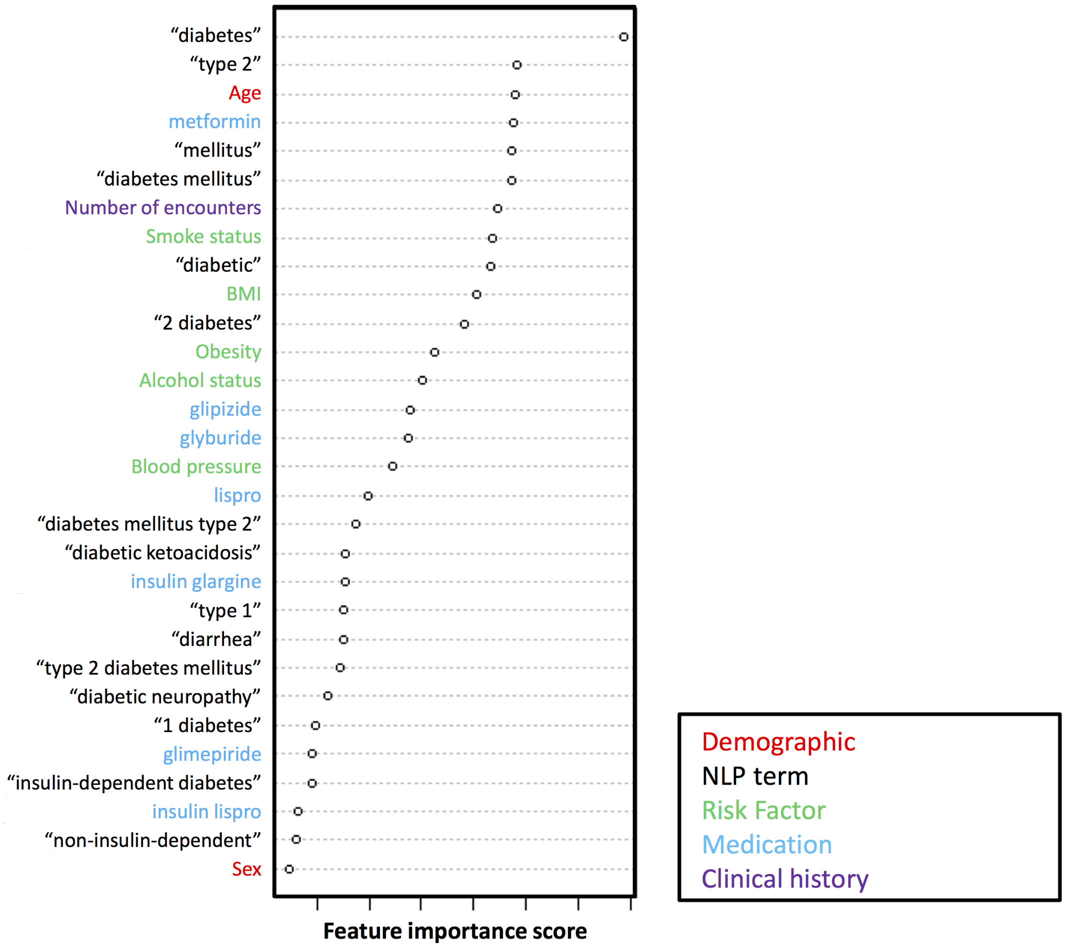
Figure 4. The multidimensional scaling (MDS) plots of the training result. This analysis was aimed at detecting meaningful underlying dimensions, for example, 1 and 2, which allow the explanation of the observed similarities (distances) between the investigated subjects. The axes of the MDS plots represent no real sizes and thus were marked as dimension 1 and dimension 2 without units. The red dots and blue triangles, indicating the positive and negative samples, were clearly separated. The "false positives" are circled in the plot. Chart reviews showed that these were notes with a genuine diagnosis of diabetes mellitus.



Diabetes Mellitus Discriminant Variables

A total of 100 DM discriminant features were retained in the final model, including demographics (n=2), risk factors (n=5), clinical history (n=1), medications (n=20), and NLP-extracted clinical terms (n=72; Multimedia Appendix 1). Figure 5 shows the top 30 features ranked by their importance in the model. The importance of each feature was rated according to the mean decrease in algorithm accuracy scaled by standard deviation after randomly permuting the variable values. A higher mean decrease in accuracy (node impurities from splitting on the variables; specifically, the node impurity is measured by the Gini index) corresponds to greater importance of the feature [40]. Among the top 30 features, "diabetes" and "type 2," which directly indicate DM, were the top 2 features, followed by age, an important predictor of DM [41,42], and then "metformin," a first-line antidiabetic drug. The remaining important discriminant features were high blood pressure, cigarette smoking, history of alcohol use, BMI, and "obesity."

Figure 5. List of the top 30 clinical variables included in the diabetes mellitus natural language processing (NLP)–based model. BMI: body mass index.



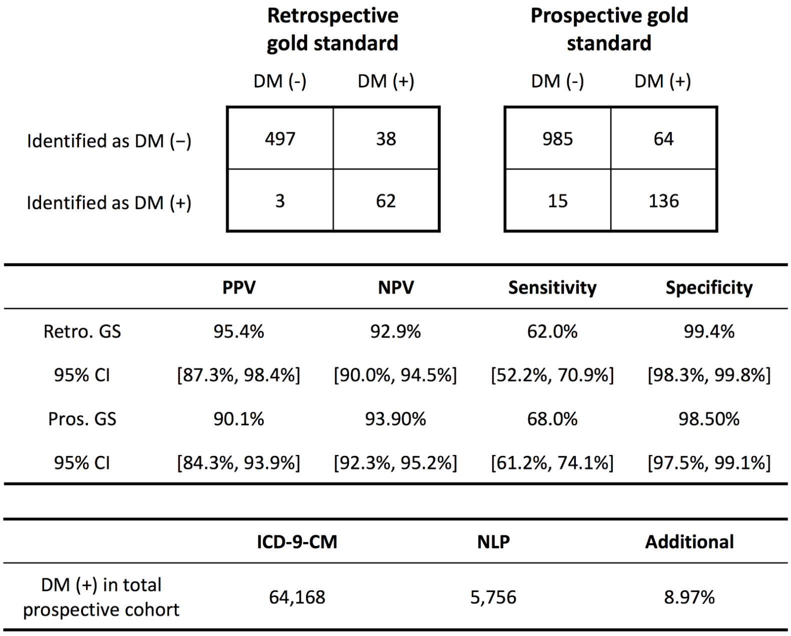
Patient Classification Cutoff Point Determination

The decision tree–based classification scores were evaluated to determine a cutoff point that allows maximal sensitivity with a ≥90% PPV (Multimedia Appendix 2). With this cutoff value (set as .618), the continuous classification scoring outputs were converted to reach a binary decision to identify genuine DM cases.

Retrospective Blind Testing

As shown in Figure 6, in the retrospective blind testing, our NLP-based analysis achieved a 95.4% (62/65) PPV, 62.0% (62/100) sensitivity, 99.4% (497/500) specificity, and 92.9% NPV (497/535). The blind testing results indicate that the knowledge acquired from some hospital facilities could be leveraged to allow prediction in others (eg, learning transfer) [43].

Figure 6. Performance evaluation of the proposed case finding algorithm. Top: the contingency tables on blind test and prospective gold standard datasets. Middle: the positive predictive value (PPV), negative predictive value (NPV), sensitivity, and specificity of the validation based on the retrospective blind testing subcohort and prospective cohort. Bottom: the prospective case finding results in the total population. DM: diabetes mellitus; GS: gold standard; ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification; NLP: natural language processing.



	Retrospective gold standard		Prospective gold standard	
	DM (-)	DM (+)	DM (-)	DM (+)
Identified as DM (−)	497	38	985	64
Identified as DM (+)	3	62	15	136

	PPV	NPV	Sensitivity	Specificity
Retro. GS	95.4%	92.9%	62.0%	99.4%
95% CI	[87.3%, 98.4%]	[90.0%, 94.5%]	[52.2%, 70.9%]	[98.3%, 99.8%]
Pros. GS	90.1%	93.90%	68.0%	98.50%
95% CI	[84.3%, 93.9%]	[92.3%, 95.2%]	[61.2%, 74.1%]	[97.5%, 99.1%]

	ICD-9-CM	NLP	Additional
DM (+) in total prospective cohort	64,168	5,756	8.97%

Prospective Validation

The prospective performance of the algorithm was explored by chart review over a gold standard dataset consisting of randomly selected 200 patients with DM and 1000 patients without DM in the uncodified subcohort (Figure 2). The PPV was 90.1% (136/151), which was within the 95% CI of the retrospective blind testing PPV (87.3%-98.4%). The sensitivity was 68.0% (136/200). The specificity, NPV, and area under ROC curve were 98.50% (985/1000), 93.90% (985/1049), and .929, respectively (Figure 6).

The algorithm was deployed to allow real-time DM case finding on the entire prospective cohort. A total of 64,168 patients with DM were identified from codified DM diagnosis, while our NLP-based algorithm identified an additional 5756 patients, resulting in an 8.97% (5756/64,168) increase in the total patients with DM during the study period. To further explore the case finding accuracy, we randomly selected 200 samples from the 5756 samples. Manual review showed that of the 200 samples there were 183 DM cases and 17 normal patients, resulting in an accuracy of 91.5% (183/200). Such accuracy was above the predetermined PPV (90%) in the calibration phase and was within the 95% CI of the retrospective blind testing PPV (87.3%-98.4%). The consistency of performance shows that it is reasonable to use the results obtained on smaller samples to reflect the performance of the algorithm on a large population.

Temporal Comparison

The time point when a patient's first DM diagnosis was identified by ICD codes was evaluated and compared with the time point when the DM was identified by NLP case finding algorithm. Out of 21,720 patients with DM identified by both methods, 6616 patients (6616/21,720, 30.46%) were identified by the NLP-based algorithm before a DM ICD code was noted in the medical record (mean time difference = 48 days). In particular, 19.86% (1314/6616) of patients were identified by NLP case finding 3 months or more before they were identified by a DM ICD code (Multimedia Appendix 3).

Discussion

Principal Findings

To the best of our knowledge, this is the first online deployment of a real-time NLP-based case finding method for DM, using both patients' structured (eg, codified diagnosis) and unstructured (free text) clinical histories from a statewide EMR database. Consistent with our hypothesis, during a 1-year period (from July 1, 2013, to June 30, 2014), our algorithm identified

XSL•FO
RenderX

5756 additional patients with DM (an 8.97% increase in the total patients with DM) who were otherwise left undiagnosed when only code-based case finding was applied. Our finding indicates that the proportion of false negatives decreased using the NLP-based approach compared with the existing ICD-based approach (*P*<.01). Many patients with DM who were misclassified as patients without DM by the code-based case finding were correctly identified by our NLP text searching algorithm, resulting in a more complete ascertainment of DM diagnoses.

There exist several reasons why patients with diagnosed DM may have not been associated with a DM diagnosis code. Among the uncodified DM patients we identified, 30% had DM noted as secondary, discharged, or other types of diagnosis and 63% had a history of diabetes in clinical records. A possible reason for missing diagnostic codes in those cases might be that if a patient was admitted to the hospital owing to more acute or life-threatening clinical conditions, information related to DM was overlooked when ICD coding was conducted. Therefore, there is a strong need for enhancing the current ICD coding practice in hospitals and other health care facilities in the state of Maine to ensure that all DM diagnoses noted in the patients' medical records are coded.

Strengths and Limitations

Although several standardized coding systems (eg, ICD, Logical Observation Identifiers Names and Codes) have been used to record diagnoses, procedures, laboratory tests, and medications associated with each patient encounter, a large amount of information related to patients' clinical histories were also available in the form of unstructured free text in EMRs. In addition to the terms directly describing DM (eg, "diabetic," "type 1," "diabetes mellitus"), our NLP algorithm was able to obtain more complete medical histories based on information about risk factors and medications available from clinical notes. A range of conventional DM risk markers (eg, age, smoking, BMI, and blood pressure) [42,44-46], emerging risk markers (eg, overweight) [47], and antidiabetic drugs (eg, metformin) were identified and used to enhance DM case detection. In particular, metformin, the first-line medication for type 2 diabetes, appeared to be the most important drug in our feature selection process. These findings indicate that our algorithm effectively incorporated a variety of clinically relevant features, leading to a significant improvement in DM case finding in the population of the state of Maine.

Another strength of our NLP case finding algorithm is the ability to find uncodified DM cases before the assignment of ICD-9-CM codes. The proposed DM case finding methodology used NLP algorithm in parallel with ICD-9-CM codes. In the prospective study, 69,924 patients with DM were identified. Among those 69,924 patients, 21,720 patients were able to be identified by both methods. That is, there were 21,720 DM patients having clinical notes that indicated they had DM. 30.46% (6616/21,720) of those patients had such clinical notes associated with an encounter earlier than the assignment of a DM diagnosis code, while 69.54% (15,104/21,720) of those patients had such clinical notes during the same encounter when a DM diagnosis code was given. Compared with using ICD-9-CM codes alone, the NLP algorithm was able to identify 30.46% (6616/21,720) of patients with DM at an earlier encounter, giving a mean time difference of 48 days. More importantly, a significant proportion of these patients (1314/6616, 19.86%) were identified 3 months or more before a DM diagnosis code was noted. For those patients, this time period is sufficient to initiate aggressive lifestyle interventions that have a long-term impact to delay progression and prevent complications of diabetes [48]. Thus, this early detection capability is clearly an advantage of our DM NLP algorithm such that these high-risk individuals can be selected for timely initiation of targeted prevention, care, and treatment.

We noted that there are some limitations in our study. First, although the use of statistical learning improved the performance of the case finding algorithm, it has inevitable misclassification errors. There were a couple of DM cases located close to the "borderline," that is, the cutoff point for the algorithm to differentiate between DM cases and normal samples. The DM cases with outputs closed to the cutoff point for the algorithm were those who were susceptible to misclassification errors, compromising false negatives. DM cases at borderline represented DM patients with incomplete DM feature profile, that is, patients having no DM-related risk factors or medication records but having clinical notes confirming DM diagnosis, or patients having no DM-related risk factors or clinical notes but having medication records. Such incomplete profiles could mislead the algorithm. Second, the relatively small sample size of the "gold standard" dataset introduced the possibility that some relatively rare clinical phenotypes of DM—where clinicians documented diabetes in a nonstandard way—might not be accounted for during model training. Third, we were unable to identify whether the patients with DM found by the NLP algorithm were those with newly diagnosed DM or those with a long-standing diagnosis. Fourth, we acknowledge our case finding method's limitation that it depends on the physician's diagnosis of the disease and the documentation quality in clinical notes. Finally, the model was developed on the patient data in the state of Maine. Extra risk factors such as sociodemographic factors may need to be considered for adjustment purpose when this learning is transferred and applied to other geographic regions.

A Web-based Identification Tool

Our NLP algorithm has been deployed online through integration into the Maine State HIE workflow, currently allowing real-time statewide identification of patients with uncodified DM. It provides doctors, hospitals, and other providers in the state HealthInfoNet network with an effective online utility to achieve a more complete assessment of the DM burden in their location. Incorporating the DM case finding algorithm with the existing health care system makes the best use of information available in EMRs. Together with the previously successful integration of our other NLP case finding algorithms, including that for congestive heart failure [14], there is a strong potential to expand the application of this novel method to enhance case finding for other diseases in Maine and other states in the United States and in other countries.

XSL•FO
RenderX

Conclusions

Our NLP-based DM case finding algorithm was developed and validated on a population-based dataset in the state of Maine. The results strongly support our hypothesis that the NLP-based algorithm could identify additional patients with DM to complement the existing ICD-code–based case finding method. Online real-time integration of our DM case finding algorithm into the Maine HIE workflow can enhance DM case detection and facilitate efforts toward timely initiation of targeted management and care for patients with DM. From the patient's perspective, many patients with DM across the state of Maine, who were not identified from ICD codified diagnosis, would benefit from information we provide by being able to take initiatives to seek care and plan their personal strategies to monitor and control their diabetes status. In this regard, our online real-time DM case finding utility not only benefits all stakeholders including payers, providers, and policy makers in the Maine health care system, but also serves as a demonstrative Web-based project for future application to improve DM case finding for targeted care and treatment in other states and countries, making a contribution to alleviate the DM burden.

Authors' Contributions

LZ, YW, SH, BJ, ADN, MJB, DJF, TF, KZ, XZ, YML, CZ, DD, YY, GZ, and DBM contributed to the analysis and interpretation of data; DSC and STA contributed to the data collection; and AYS, FS, KGS, EW, and XBL contributed to conception or design of the work. LZ, YW, and SH drafted the manuscript; LZ, YW, SH, BJ, ADN, MJB, DJF, TF, KZ, XZ, YML, CZ, DD, YY, GZ, DBM, DSC, STA, AYS, FS, KGS, EW, and XBL critically revised the manuscript.

Conflicts of Interest

The authors have the following interests: KGS, EW, and XBL are cofounders and equity holders of HBI Solutions, Inc, which is currently developing predictive analytics solutions for health care organizations. From the departments of pediatrics, surgery, and cardiothoracic surgery, Stanford University School of Medicine, Stanford, California, Zhejiang University School of Medicine and School of Management, Pingjin Hospital Heart Center, and Tsinghua University School of Electrical Engineering, LZ, YW, SH, AYS, KZ, XZ, YML, YY, GZ, DBM, KGS, and XBL conducted this research as part of a personal outside consulting arrangement with HBI Solutions, Inc. The research and research results are not, in any way, associated with these institutions. This does not alter the authors' adherence to all the journal policies on sharing data and materials, as detailed online in the guide for authors.

Multimedia Appendix 1

A list of 100 discriminant features used by the final model as well as the feature importance, and a list of 742 natural language processing terms used by the initial modeling process.

[PDF File (Adobe PDF File), 55KB - medinform_v4i4e37_app1.pdf]

Multimedia Appendix 2

Determination of the cutoff point of the subject classification probabilities. Top: the receiver operating characteristic curve and the line determined by the prevalence and 90% positive predictive value were intersected at the cutoff point. Bottom: their intersection (dashed rectangle) is magnified and indicated by the circle.

[PDF File (Adobe PDF File), 73KB - medinform_v4i4e37_app2.pdf]

Multimedia Appendix 3

The distribution of patients by the time intervals between natural language processing–based diabetes mellitus identification and codified diagnosis.

[PDF File (Adobe PDF File), 50KB - medinform_v4i4e37_app3.pdf]

References

1. Murray CJ, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012 Dec 15;380(9859):2197-2223. [doi: 10.1016/S0140-6736(12)61689-4] [Medline: 23245608]
2. Dörhöfer L, Lammert A, Krane V, Gorski M, Banas B, Wanner C, et al. Study design of DIACORE (DIAbetes COhoRtE) - a cohort study of patients with diabetes mellitus type 2. BMC Med Genet 2013;14:25 [FREE Full text] [doi: 10.1186/1471-2350-14-25] [Medline: 23409726]
3. Centers for Disease Control and Prevention. Atlanta, GA: US Department of Health and Human Services; 2011. National Diabetes Fact Sheet: National estimates and general information on diabetes and prediabetes in the United States, 2011 URL: https://www.cdc.gov/diabetes/pubs/pdf/ndfs_2011.pdf [accessed 2016-07-03] [WebCite Cache ID 6io6Wx73s]

XSL•FO
RenderX

4. Centers for Disease Control and Prevention (CDC). Atlanta, GA: US Department of Health and Human Services; 2014. National Diabetes Statistics Report: Etimates of Diabetes and Its Burden in the United States, 2014 URL: https://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-report-web.pdf [accessed 2016-07-04] [WebCite Cache ID 6io7cyhpe]
5. Nathan DM, Cleary PA, Backlund JC, Genuth SM, Lachin JM, Orchard TJ, Diabetes ControlComplications Trial/Epidemiology of Diabetes InterventionsComplications (DCCT/EDIC) Study Research Group. Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes. N Engl J Med 2005 Dec 22;353(25):2643-2653 [FREE Full text] [doi: 10.1056/NEJMoa052187] [Medline: 16371630]
6. Gong Q, Gregg EW, Wang J, An Y, Zhang P, Yang W, et al. Long-term effects of a randomised trial of a 6-year lifestyle intervention in impaired glucose tolerance on diabetes-related microvascular complications: the China Da Qing Diabetes Prevention Outcome Study. Diabetologia 2011 Feb;54(2):300-307. [doi: 10.1007/s00125-010-1948-9] [Medline: 21046360]
7. Tchobroutsky G. Relation of diabetic control to development of microvascular complications. Diabetologia 1978 Sep;15(3):143-152. [Medline: 359393]
8. Engerman R, Bloodworth Jr JM, Nelson S. Relationship of microvascular disease in diabetes to metabolic control. Diabetes 1977 Aug;26(8):760-769. [Medline: 885298]
9. Wei W, Leibson CL, Ransom JE, Kho AN, Chute CG. The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects. Int J Med Inform 2013 Apr;82(4):239-247 [FREE Full text] [doi: 10.1016/j.ijmedinf.2012.05.015] [Medline: 22762862]
10. Khokhar B, Jette N, Metcalfe A, Cunningham CT, Quan H, Kaplan GG, et al. Systematic review of validated case definitions for diabetes in ICD-9-coded and ICD-10-coded data in adult populations. BMJ Open 2016 Aug;6(8):e009952 [FREE Full text] [doi: 10.1136/bmjopen-2015-009952] [Medline: 27496226]
11. Vassar M, Holzmann M. The retrospective chart review: important methodological considerations. J Educ Eval Health Prof 2013;10:12 [FREE Full text] [doi: 10.3352/jeehp.2013.10.12] [Medline: 24324853]
12. Shine D, Sundaram P, Torres DM, Johnstone B, Jaeger J, Sanguliano B. Can computerized cost data substitute for chart review? J Healthc Qual 2002;24(6):26-33. [Medline: 12432860]
13. 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]
14. Wang Y, Luo J, Hao S, Xu H, Shin AY, Jin B, et al. NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records. Int J Med Inform 2015 Dec;84(12):1039-1047. [doi: 10.1016/j.ijmedinf.2015.06.007] [Medline: 26254876]
15. Pakhomov SV, Buntrock J, Chute CG. Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier. J Biomed Inform 2005 Apr;38(2):145-153 [FREE Full text] [doi: 10.1016/j.jbi.2004.11.016] [Medline: 15797003]
16. Carrell DS, Halgrim S, Tran D, Buist DS, Chubak J, Chapman WW, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. Am J Epidemiol 2014 Mar 15;179(6):749-758 [FREE Full text] [doi: 10.1093/aje/kwt441] [Medline: 24488511]
17. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff (Millwood) 2014 Jul;33(7):1163-1170. [doi: 10.1377/hlthaff.2014.0053] [Medline: 25006142]
18. Jacob JA. On the Road to Interoperability, Public and Private Organizations Work to Connect Health Care Data. JAMA 2015 Sep;314(12):1213-1215. [doi: 10.1001/jama.2015.5930] [Medline: 26393833]
19. Ng K, Ghoting A, Steinhubl SR, Stewart WF, Malin B, Sun J. PARAMO: a PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. J Biomed Inform 2014 Apr;48:160-170 [FREE Full text] [doi: 10.1016/j.jbi.2013.12.012] [Medline: 24370496]
20. Forrest CB, Margolis PA, Bailey LC, Marsolo K, Del Beccaro MA, Finkelstein JA, et al. PEDSnet: a National Pediatric Learning Health System. J Am Med Inform Assoc 2014;21(4):602-606 [FREE Full text] [doi: 10.1136/amiajnl-2014-002743] [Medline: 24821737]
21. Sittig DF, Wright A. What makes an EHR “open” or interoperable? J Am Med Inform Assoc 2015 Sep;22(5):1099-1101 [FREE Full text] [doi: 10.1093/jamia/ocv060] [Medline: 26078411]
22. Hinfonet. HealthInfoNet 2016 URL: http://hinfonet.org/ [accessed 2016-09-30] [WebCite Cache ID 6kvKLP6bc]
23. Hao S, Jin BO, Shin AY, Zhao Y, Zhu C, Li Z, et al. Risk prediction of emergency department revisit 30 days post discharge: a prospective study. PLoS One 2014 Nov;9(11):e112944 [FREE Full text] [doi: 10.1371/journal.pone.0112944] [Medline: 25393305]
24. Holt TA, Gunnarsson CL, Cload PA, Ross SD. Identification of undiagnosed diabetes and quality of diabetes care in the United States: cross-sectional study of 11.5 million primary care electronic records. CMAJ Open 2014 Oct;2(4):E248-E255 [FREE Full text] [doi: 10.9778/cmajo.20130095] [Medline: 25485250]
25. Schulze MB, Hoffmann K, Boeing H, Linseisen J, Rohrmann S, Möhlig M, et al. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. Diabetes Care 2007 Mar;30(3):510-515. [doi: 10.2337/dc06-2089] [Medline: 17327313]

XSL•FO
RenderX

26. Liu M, Pan C, Jin M. A Chinese diabetes risk score for screening of undiagnosed diabetes and abnormal glucose tolerance. Diabetes Technol Ther 2011 May;13(5):501-507. [doi: 10.1089/dia.2010.0106] [Medline: 21406016]
27. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med 2011 Sep;9:103 [FREE Full text] [doi: 10.1186/1741-7015-9-103] [Medline: 21902820]
28. Balkau B, Lange C, Fezeu L, Tichet J, de Lauzon-Guillain B, Czernichow S, et al. Predicting diabetes: clinical, biological, and genetic approaches: data from the Epidemiological Study on the Insulin Resistance Syndrome (DESIR). Diabetes Care 2008 Oct;31(10):2056-2061 [FREE Full text] [doi: 10.2337/dc08-0368] [Medline: 18689695]
29. Aekplakorn W, Bunnag P, Woodward M, Sritara P, Cheepudomwit S, Yamwong S, et al. A risk score for predicting incident diabetes in the Thai population. Diabetes Care 2006 Aug;29(8):1872-1877. [doi: 10.2337/dc05-2141] [Medline: 16873795]
30. Kim TY, Hardiker N, Coenen A. Inter-terminology mapping of nursing problems. J Biomed Inform 2014 Jun;49:213-220 [FREE Full text] [doi: 10.1016/j.jbi.2014.03.001] [Medline: 24632297]
31. Nadkarni PM, Darer JA. Migrating existing clinical content from ICD-9 to SNOMED. J Am Med Inform Assoc 2010;17(5):602-607 [FREE Full text] [doi: 10.1136/jamia.2009.001057] [Medline: 20819871]
32. Lipscomb CE. Medical Subject Headings (MeSH). Bull Med Libr Assoc 2000 Jul;88(3):265-266 [FREE Full text] [Medline: 10928714]
33. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: the R Foundation for Statistical Computing; 2015.
34. WHO Expert Consultation. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. Lancet 2004 Jan 10;363(9403):157-163. [doi: 10.1016/S0140-6736(03)15268-3] [Medline: 14726171]
35. Chapman WW, Hillert D, Velupillai S, Kvist M, Skeppstedt M, Chapman BE, et al. Extending the NegEx lexicon for multiple languages. Stud Health Technol Inform 2013;192:677-681 [FREE Full text] [Medline: 23920642]
36. English-for-students. Family Vocabulary Word List 2015 URL: http://www.english-for-students.com/Family-Vocabulary.html [accessed 2016-07-05] [WebCite Cache ID 6io7Hc6c3]
37. Doan S, Maehara CK, Chaparro JD, Lu S, Liu R, Graham A, Pediatric Emergency Medicine Kawasaki Disease Research Group. Building a Natural Language Processing Tool to Identify Patients With High Clinical Suspicion for Kawasaki Disease from Emergency Department Notes. Acad Emerg Med 2016 May;23(5):628-636. [doi: 10.1111/acem.12925] [Medline: 26826020]
38. Breiman L. Random forests. Machine Learning 2001;45(1):5-32. [doi: 10.1023/A:1010933404324]
39. Love TJ, Cai T, Karlson EW. Validation of psoriatic arthritis diagnoses in electronic medical records using natural language processing. Semin Arthritis Rheum 2011 Apr;40(5):413-420 [FREE Full text] [doi: 10.1016/j.semarthrit.2010.05.002] [Medline: 20701955]
40. Stat.berkeley. Random Forests URL: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm [accessed 2016-10-01] [WebCite Cache ID 6kvL2QzSq]
41. Nayak BS, Sobrian A, Latiff K, Pope D, Rampersad A, Lourenço K, et al. The association of age, gender, ethnicity, family history, obesity and hypertension with type 2 diabetes mellitus in Trinidad. Diabetes Metab Syndr 2014;8(2):91-95. [doi: 10.1016/j.dsx.2014.04.018] [Medline: 24907173]
42. Ding D, Chong S, Jalaludin B, Comino E, Bauman AE. Risk factors of incident type 2-diabetes mellitus over a 3-year follow-up: Results from a large Australian sample. Diabetes Res Clin Pract 2015 May;108(2):306-315. [doi: 10.1016/j.diabres.2015.02.002] [Medline: 25737033]
43. Wiens J, Guttag J, Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. J Am Med Inform Assoc 2014;21(4):699-706 [FREE Full text] [doi: 10.1136/amiajnl-2013-002162] [Medline: 24481703]
44. Jerant A, Franks P. Body mass index, diabetes, hypertension, and short-term mortality: a population-based observational study, 2000-2006. J Am Board Fam Med 2012;25(4):422-431 [FREE Full text] [doi: 10.3122/jabfm.2012.04.110289] [Medline: 22773710]
45. Condliffe S, Link CR, Parasuraman S, Pollack MF. The effects of hypertension and obesity on total health-care expenditures of diabetes patients in the United States. Applied Economics Letters 2013 May;20(7):649-652. [doi: 10.1080/13504851.2012.727966]
46. Araneta MR, Kanaya AM, Hsu WC, Chang HK, Grandinetti A, Boyko EJ, et al. Optimum BMI cut points to screen asian americans for type 2 diabetes. Diabetes Care 2015 May;38(5):814-820 [FREE Full text] [doi: 10.2337/dc14-2071] [Medline: 25665815]
47. American Diabetes Association. Classification and diagnosis of diabetes. Sec. 2. In Standards of Medical Care in Diabetes-2016. Diabetes Care 2016 Jan;39(Suppl 1):S13-S22. [doi: 10.2337/dc16-S005] [Medline: 26696675]
48. Schellenberg ES, Dryden DM, Vandermeer B, Ha C, Korownyk C. Lifestyle interventions for patients with and at risk for type 2 diabetes: a systematic review and meta-analysis. Ann Intern Med 2013 Oct 15;159(8):543-551. [doi: 10.7326/0003-4819-159-8-201310150-00007] [Medline: 24126648]

XSL•FO
RenderX

Abbreviations

BMI: body mass index
DM: diabetes mellitus
EMR: electronic medical record
HIE: health information exchange
ICD: International Classification of Diseases
ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification
MDS: multidimensional scaling
MeSH: Medical Subject Headings
NLM: US National Library of Medicine
NLP: natural language processing
NPV: negative predictive value
PPV: positive predictive value
ROC: receiver operating characteristic
SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms

Edited by G Eysenbach; submitted 07.07.16; peer-reviewed by F Pourmalek, J Harrison; comments to author 02.08.16; revised version received 01.10.16; accepted 12.10.16; published 11.11.16.

Please cite as:
Zheng L, Wang Y, Hao S, Shin AY, Jin B, Ngo AD, Jackson-Browne MS, Feller DJ, Fu T, Zhang K, Zhou X, Zhu C, Dai D, Yu Y, Zheng G, Li YM, McElhinney DB, Culver DS, Alfreds ST, Stearns F, Sylvester KG, Widen E, Ling XB
Web-based Real-Time Case Finding for the Population Health Management of Patients With Diabetes Mellitus: A Prospective Validation of the Natural Language Processing–Based Algorithm With Statewide Electronic Medical Records
JMIR Med Inform 2016;4(4):e37
URL: http://medinform.jmir.org/2016/4/e37/
doi:10.2196/medinform.6328
PMID:27836816

©Le Zheng, Yue Wang, Shiying Hao, Andrew Y Shin, Bo Jin, Anh D Ngo, Medina S Jackson-Browne, Daniel J Feller, Tianyun Fu, Karena Zhang, Xin Zhou, Chunqing Zhu, Dorothy Dai, Yunxian Yu, Gang Zheng, Yu-Ming Li, Doff B McElhinney, Devore S Culver, Shaun T Alfreds, Frank Stearns, Karl G Sylvester, Eric Widen, Xuefeng Bruce Ling. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 11.11.2016. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Consumers’ Use of UMLS Concepts on Social Media: Diabetes-Related Textual Data Analysis in Blog and Social Q&A Sites

Min Sook Park[1*], PhD; Zhe He[1,2*], PhD; Zhiwei Chen[3], BEng; Sanghee Oh[1], PhD; Jiang Bian[4], PhD

[1]School of Information, Florida State University, Tallahassee, FL, United States

[2]Institute for Successful Longevity, Florida State University, Tallahassee, FL, United States

[3]Department of Computer Science, Florida State University, Tallahassee, FL, United States

[4]Department of Health Outcomes and Policy, University of Florida, Gainesville, FL, United States

[*]these authors contributed equally

Corresponding Author:
Zhe He, PhD
School of Information
Florida State University
Louis Shores Building
142 Collegiate Loop
Tallahassee, FL, 32306
United States
Phone: 1 850 644 5775
Fax: 1 850 644 9763
Email: zhe.he@cci.fsu.edu

Abstract

Background: The widely known terminology gap between health professionals and health consumers hinders effective information seeking for consumers.

Objective: The aim of this study was to better understand consumers’ usage of medical concepts by evaluating the coverage of concepts and semantic types of the Unified Medical Language System (UMLS) on diabetes-related postings in 2 types of social media: blogs and social question and answer (Q&A).

Methods: We collected 2 types of social media data: (1) a total of 3711 blogs tagged with “diabetes” on Tumblr posted between February and October 2015; and (2) a total of 58,422 questions and associated answers posted between 2009 and 2014 in the diabetes category of Yahoo! Answers. We analyzed the datasets using a widely adopted biomedical text processing framework Apache cTAKES and its extension YTEX. First, we applied the named entity recognition (NER) method implemented in YTEX to identify UMLS concepts in the datasets. We then analyzed the coverage and the popularity of concepts in the UMLS source vocabularies across the 2 datasets (ie, blogs and social Q&A). Further, we conducted a concept-level comparative coverage analysis between SNOMED Clinical Terms (SNOMED CT) and Open-Access Collaborative Consumer Health Vocabulary (OAC CHV)—the top 2 UMLS source vocabularies that have the most coverage on our datasets. We also analyzed the UMLS semantic types that were frequently observed in our datasets.

Results: We identified 2415 UMLS concepts from blog postings, 6452 UMLS concepts from social Q&A questions, and 10,378 UMLS concepts from the answers. The medical concepts identified in the blogs can be covered by 56 source vocabularies in the UMLS, while those in questions and answers can be covered by 58 source vocabularies. SNOMED CT was the dominant vocabulary in terms of coverage across all the datasets, ranging from 84.9% to 95.9%. It was followed by OAC CHV (between 73.5% and 80.0%) and Metathesaurus Names (MTH) (between 55.7% and 73.5%). All of the social media datasets shared frequent semantic types such as “Amino Acid, Peptide, or Protein,” “Body Part, Organ, or Organ Component,” and “Disease or Syndrome.”

Conclusions: Although the 3 social media datasets vary greatly in size, they exhibited similar conceptual coverage among UMLS source vocabularies and the identified concepts showed similar semantic type distributions. As such, concepts that are both frequently used by consumers and also found in professional vocabularies such as SNOMED CT can be suggested to OAC CHV to improve its coverage.

XSL•FO
RenderX

(JMIR Med Inform 2016;4(4):e41) doi:10.2196/medinform.5748

KEYWORDS

controlled vocabulary; consumer health vocabulary; concept coverage

Introduction

Background

There is a widely known language gap between health consumers and health care professionals [1-3]. This gap may hinder effective communication between the 2 groups [4-7]; thus, impacting consumers' health information seeking [3,8,9] and subsequent decision making regarding their health issues [10]. To assess the gap, Roberts and Demner-Fushman [11] used a variety of natural language processing (NLP) techniques to analyze the difference between health questions asked by consumers and health professionals in different online question and answer (Q&A) sites (eg, Yahoo! Answers, and WebMD). They found that consumer questions tend to contain more misspelled medical terms, have longer background information, and resemble open-domain language more closely than texts written by professionals. One major aspect of the gap is the difference in medical vocabulary used by consumers and health professionals. Zeng and colleagues [12] observed that when searching online health information, using only consumer terms leads to poor information retrieval results. Plovnick and Zeng [13] later reformulated consumers' health queries with professional terminology and about 40% of reformulated queries yielded better search performance.

To bridge the vocabulary gap between health professionals and consumers, early researchers have collected and analyzed diverse textual data generated by consumers to identify medical terms used by consumers. Brennan and Aronson [14] used the MetaMap tool to extract salient concepts in nursing vocabularies from consumers' email messages. Smith and collegues [15] also used MetaMap to successfully identify the Unified Medical Language System (UMLS) concepts used by consumers in their email messages submitted to University of Pittsburg Cancer Institute's Cancer Information and Referral Service. These studies aimed to bridge the vocabulary gap between health professionals and consumers by identifying frequently-used consumer health terms that are relevant in developing consumer-oriented health information applications and linking free text to complex clinical knowledge resources. These *ad hoc* studies represent early efforts in bridging the vocabulary gap.

A controlled vocabulary is "an organized arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching[16]." In an effort to formalize consumer vocabulary for various applications, a controlled vocabulary called Open-Access Collaborative Consumer Health Vocabulary ("OAC CHV," "CHV" for short) was recently developed as a collection of expressions and concepts that are commonly used by ordinary health information users [17]. Moreover, CHV has been integrated in the largest medical terminological system–the UMLS, which has mapped terms from different source vocabularies with the same meaning into the same concept by the United States National Library of Medicine (NLM). As such, consumer terms are connected to their corresponding professional terms in professional vocabularies such as SNOMED Clinical Terms (SNOMED CT). With CHV in the UMLS, one can translate a sentence with consumer terms to a sentence with professional terms in an automated fashion.

Domain coverage—the extent to which a controlled vocabulary covers the intended domain—is one of the most desired properties for a controlled vocabulary [18]. The usability and the overall structure of a controlled vocabulary heavily rely upon its coverage [19]. Traditionally, controlled vocabulary development takes a top-down approach, which reflects a group of experts' knowledge in the respective subject matter [20,21]. For the development of CHV, however, a bottom-up approach was taken, emphasizing 2 fundamental properties: (1) CHV should capture actual consumers' terms and expressions that reflect their health information needs, and (2) the expressions should be familiar to and used by consumers [17].

To keep up with continuous evolution of medical knowledge, CHV needs to be updated and maintained by incorporating new, consumer-provided terms and expressions [17,22-24]. Existing studies have shown promising results in discovering consumer terms for CHV from social media, in particular. Vydiswaran et al [7] applied a pattern-based text mining approach to identify pairs of consumer and professional terms from Wikipedia. Hicks et al [25] analyzed consumer messages exchanged in Twitter in order to evaluate terms related to gender identification on intake forms. Doing-Harris and Zeng-Treitler [24] developed a computer assisted CHV update system, which can automatically identify prospective terms from social media. Identifying terms used by consumers in consumer-generated text in aggregate fashion can account for the variability of lay health language. These terms can be used to refine and enrich CHV [17].

Consumers, however, may also learn and use professional terms [17,24,26]. In this sense, medical terms that are familiar to consumers and are already established in other controlled vocabularies could be used to improve the coverage of CHV. Term reuse is a principle and best practice in ontology/terminology development as it promises to support the semantic interoperability and to reduce engineering costs [27]. Researchers have previously developed semi-automated methods to facilitate systematic term reuse. He et al [28] developed a topological-pattern-based method to identify terms from UMLS source vocabularies to enrich SNOMED CT [28,29] and National Cancer Institute Thesaurus (NCIt) [30].

However, this method cannot be directly applied to CHV, because it does not have hierarchical relationships (eg. parent-child relationship) that are necessary to construct topological patterns [28-30]. Recently, Chandar et al [31] introduced a similarity-based term recommendation method that represents n-grams extracted from the free-text eligibility

XSL•FO
RenderX

criteria of clinical trials as a set of linguistic and contextual features. SNOMED CT terms are clustered with K-means clustering. The new terms are ordered by their distance to the nearest cluster centroid, representing their similarity to existing SNOMED CT terms. This method performed well on the corpus of free-text clinical study eligibility criteria, because they are mostly short and partial sentences written by health professionals with fruitful medical terms and little noise. It has yet to be tested on free-form consumer text that typically contains lengthy sentences and lay terms.

Most previous studies concerning CHV development concentrated on the identification of new terms used by consumers [17,22-24]. To the best of our knowledge, no prior studies have conducted in-depth assessment of the coverage and popularity of medical concepts in user-generated documents on social media. In this respect, there is a need to understand consumers' use of terms in existing controlled vocabularies, and to perceive if there is the potential to improve CHV by incorporating health-related concepts used by consumers that are covered by professional vocabularies. In this study, therefore, we performed such an analysis in order to assess consumers' use of medical concepts on social media postings pertaining to health concerns and to evaluate how many popular consumer terms have been included in the existing source vocabularies of the UMLS [32].

In this study, we focus on diabetes, which is recognized as one of the most important public health problems with escalating health concerns by the World Health Organization (WHO) [33]. Diabetes caused 1.5 million deaths in 2012 alone. It is known to cause disability and an array of serious health issues such as hypertension, nephropathy, and stroke [34]. Global diabetes cases skyrocketed from 108 million in 1980 to 422 million in 2014. The number of diabetes onset will likely reach 700 million by 2025 [35]. Diabetes and its complications not only impair population health but also impose substantial economic burdens on patients, their family, and the society [33].

In this study, we collected diabetes-related consumer-generated blog postings from Tumblr and diabetes-related questions and answers from Yahoo! Answers. We carried out text mining to identify UMLS concepts from our datasets. Thus, we formulated the 2 research questions (RQs): (1) To what degree do the concepts from UMLS source vocabularies cover the concepts used by consumers describing their diabetes-related concerns on health postings of social media, especially blogs and social Q&A? Which concepts do or do not overlap? (2) To what degree are the UMLS semantic types applicable to analyzing the concepts used by consumers when describing their diabetes-related concerns in social media, especially blogs and social Q&A? Which semantic types are observed?

In the first research question, we evaluated the coverage of all of the 178 English source vocabularies of the UMLS in our 2 datasets from Tumblr and Yahoo! Answers. In the second research question, we analyzed the semantic types of the UMLS concepts identified in our datasets.

The current study mainly investigated the overlap between consumer concepts from social media and professional concepts in the UMLS. Indeed, consumers often proactively seek and share online health information on social media [36,37]. Their use of professional terms could be sophisticated covering both laypersons' expressions and medical terminologies. In fact, not only consumers but also health care professionals have actively participated in creating health postings in social media [38,39]. Their use of terms in social media, however, is likely to be more consumer/patient-centric for health education and promotion to the public. The comparative analysis of the concept coverage between consumers and professional vocabularies in social media may be helpful in understanding the scale of the phenomenon. The comparison will also help yield insights into the nature of the vocabulary gap, which will contribute to the consistent development of the CHV. The current study, in particular, could shed light on how much social media users use existing terms in UMLS source vocabularies on the web. At the same time, findings from the current study could inform the feasibility of leveraging existing UMLS source vocabularies to enrich the CHV.

The Unified Medical Language System

The UMLS, maintained by the NLM of the National Institutes of Health, is the largest biomedical terminological system. Its 2-level structure consists of Metathesaurus and Semantic Network. The UMLS Metathesaurus is "a large, multi-purpose, and multi-lingual thesaurus that contains millions of biomedical and health related concepts, their synonymous names, and their relationships" [40]. The UMLS Metathesaurus integrates more than 9.1 million terms from over 170 English source vocabularies into 3.1 million medical concepts (2015AA version). Besides English, the UMLS also contains source vocabularies in 20 other languages. The UMLS has integrated most of the well-designed and well-maintained medical terminologies such as SNOMED CT, the International Classification of Diseases 9th Revision, Clinical Modification (ICD-9-CM), NCIt, and RxNORM. SNOMED CT is the most comprehensive and precise clinical terminology in the world with over 310,000 active concepts [41]. ICD-9-CM is used primarily to encode the diagnoses and procedures for billing purposes [42]. RxNORM, on the other hand, normalizes names of all clinical drugs available on the US market and their links to many of the drug vocabularies commonly used in pharmacy management [43]. Most significantly, the terms with the same meaning are mapped to the same concept in the UMLS. Due to its native term mapping, the UMLS is a valuable resource for supporting interoperability and translation in biomedicine [32]. The NLM releases a new version of the UMLS twice a year.

The UMLS semantic types represent "a set of broad subject categories that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus" [44]. Each concept in the UMLS is assigned 1 or more semantic types. In the 2015AA version of the UMLS, there are a total of 127 semantic types, describing concepts at the levels of entity and event. Entities include physical objects such as organism, anatomical structure, and substances. Events describe activities, phenomenon, and processes. For example, the semantic type "Disease or Syndrome" categorizes a set of concepts in the UMLS that indicate "a condition which alters or interferes with a normal process, state, or activity of an organism."

XSL•FO
RenderX

Consumer Health Vocabularies and Their Use in Consumer-Oriented Health Applications

OAC CHV has been used in various health-related applications to improve patients' access to health information. Zeng et al developed a translator specifically to convert texts in electronic health records to consumer-friendly text in patient health records by replacing UMLS terms to their corresponding OAC CHV terms [45]. Many UMLS concepts have one to one match with OAC CHV concepts. All the OAC CHV concepts have predefined consumer-friendly display names. Besides OAC CHV, other proprietary consumer health vocabularies have been developed. For example, Apelon has developed a CHV and has mapped their CHV terms to corresponding clinical concepts in SNOMED RT (an earlier version of SNOMED CT, developed by College of American Pathologists), ICD-9-CM, and Physician's Current Procedural Terminology (CPT) administrative codes. The CHV of Apelon has been used in various applications, such as consumer health data entry, patient results reporting clinical note translation, and Web-based information retrieval [46]. Mayo Clinic also developed their own consumer health vocabulary, which has a rich content of disease concepts as well as genetic and non-genetic risk factors to diseases [8]. In this paper, we used OAC CHV because it is the only publicly available consumer health vocabulary that we have access to (through the UMLS).

Methods

Data Collection

2 types of social media were analyzed in the current study, namely blogs and social Q&A, as they allow consumers to generate and freely exchange health information in text format. Health-related blogs are one of the most popular social media venues for health information distribution. Bloggers typically describe their personal experiences with diseases along with their encounters with health care professionals [47]. Health care professionals also create blogs for sharing their medical knowledge and information with patients [48]. Blogs have also been widely used for health promotion and education as a collaborative tool for both consumers and health care professionals [49-51]. On the other hand, social Q&A is an online community-based Q&A service where people gain knowledge through raising questions and receiving answers from others who willingly share their knowledge, experiences, and opinions regarding a wide range of topics including health. Social Q&A is considered to be a knowledge-shaping sphere for laypeople [52]. Consumers are motivated to use social Q&A because their searches on web search engines with short queries that are not fully expressive often fail in retrieving useful information for their specific problems, while social Q&A allows them to ask questions in natural language and in full sentences [11]. For data collection, we used 2 datasets: (1) Tumblr, a popular blogging service; and (2) Yahoo! Answers, a social Q&A service in North America.

Tumblr and Yahoo! Answers were chosen for the current study due to their popularity and the convenience of using their Application Program Interfaces (APIs), which allowed us to collect data automatically from these sites. Also, both Tumblr and Yahoo! Answers do not limit the number of words in postings. As such, their users can elaborate their health concerns and information on postings with sufficient details, thereby providing us ample opportunities to extract and analyze relevant concepts from the postings.

Tumblr is one of the fastest-growing blog sites with nearly twenty-fold increase in the number of blogs from October 2012 to October 2015 [53]. It launched relatively late in the market compared to other sites such as WordPress and Blogger, but is recognized as one of the best blog sites due to its ease of setup, stylish interface design, and micro-blogging support [54,55]. It has over 227 million blogs and 37 million unique visitors as of February 2016 [53]. From Tumblr, we collected a total of 3711 English text blogs with a tag related to "diabetes" (eg, "diabetes," "diabetes mellitus," and "Type 2 diabetes") posted between February and October 2015.

Yahoo! Answers is one of the most popular social Q&A sites with approximately 5.6 million visitors per month as of February 2016 [56]. From Yahoo! Answers, we garnered a total of 58,422 questions and associated answers between 2009 and 2014 in the diabetes category of Yahoo! Answers. During data analysis, we carried out text mining with questions and answers (specifically, best answers) separately, because the information in questions and answers could be different. Questions could capture health concerns and associated problems, while answers could mainly discuss information resources intended to solve the problems. It is important to note that 1 question may have more than one answer. In this study, we limited answers to the one selected as the best answer by the questioner. The data collected from Yahoo! Answers were separated into questions and answers in the subsequent analyses.

Units of Analysis

Once we collected text data from Tumblr and Yahoo! Answers, we mined the text data for "concepts," a unit of understanding which represents a fundamental component of terminology [57] or unit of meaning in an ontology [31]. Concepts are different from "terms" in that a term refers to an entity or "physical object" written or spoken in text to represent a concept or thought [58]. In the UMLS, a term is described as a "word or collection of words comprising an expression," which indicates a class of all lexical variants (eg, "eye," "Eye," "eyes") [59]. The UMLS assigned each term an atom unique identifier (AUI) and grouped the terms with the same meaning into a concept with a concept unique identifier (CUI). We also analyzed the semantic types of the extracted concepts in order to understand the broad semantic categories of the terms that are frequently used by consumers.

Textual Data Processing

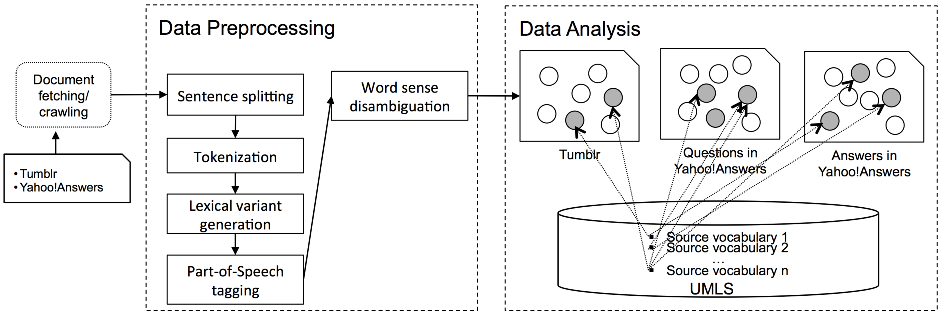
We used a widely adopted biomedical text processing framework Apache cTAKES™ [60] and its extension YTEX [61] to identify UMLS terms in our datasets. Apache cTAKES is designed as a natural language processing (NLP) system for extraction of information from the free-text data available in electronic medical records (EMRs). It provides an agile and flexible platform based on the Unstructured Information Management Architecture (UIMA) and a rich NLP library.

XSL•FO
RenderX

YTEX, a module of cTAKES, provides Word Sense Disambiguation (WSD), data mining and feature engineering functionalities. We mainly used the WSD function of YTEX to recognize the most possible UMLS concept when a term in the free text can be matched to multiple ambiguous concepts. We used the 3.2.2 release of cTAKES and YTEX with the default workflow configuration named “Aggregate Plaintext UMLS Processor.”

Figure 1 illustrates our overall analysis process. First, each document is a blog posting from Tumblr, a question or an answer from Yahoo! Answers. Each blog posting may consist of 1 or more sentences. Then, cTAKES detected and split each document into individual sentences using the sentence detector of OpenNLP [62,63], with the default configuration for English text. For each sentence, cTAKES performed tokenization with the default tokenizer of the OpenNLP, lexical variant generation using the lexical tool provided by the United States National Library of Medicine with the default configuration. Then, cTAKES performed Part-Of-Speech (POS) tagging using the POS tagger in OpenNLP with the information entropy-based model for English to generate the candidate terms for further processing. Then, YTEX matched the candidate terms with all the possible UMLS terms, which were preloaded from the MRCONSO table of the UMLS 2015AA release. We then stored the matching results to a MySQL database. For each candidate term, there may be 0, 1, or more matching UMLS terms with different semantics. To identify terms with reasonable semantics, we used YTEX to perform word sense disambiguation (WSD), in which the intrinsic information content (IC) measure is used as the semantic similarity metric with a window of 50 words as the context for WSD. The intrinsic information content is a measure of concept specificity computed from the structure of the taxonomy in a biomedical terminology and does not rely on the term frequency in the corpus. The details of the intrinsic IC measure can be found in Garla et al [64]. Finally, all the UMLS terms in each record were extracted with a UMLS CUI.

Figure 1. Conceptual framework of the study. Dots refer to concepts extracted from the dataset and gray dots refer to concepts mapped to the concepts in one of the UMLS source vocabularies.



Concept Coverage Analysis

We first analyzed the basic characteristics of the overall concept coverage across our datasets collected from Tumblr and Yahoo! Answers: (1) blog postings from Tumblr, (2) questions in Yahoo! Answers, and (3) answers in Yahoo! Answers. We then analyzed the coverage of each source vocabulary in the UMLS across the datasets. SNOMED CT and CHV are the 2 vocabularies with the highest concept coverage in our datasets. Thus, we conducted a concept coverage analysis of SNOMED CT and CHV based on our datasets. We also analyzed the semantic types of the concepts identified from our datasets.

Results

Aggregate Characteristics of the Datasets

We identified 2415 UMLS concepts from blog postings, 6452 UMLS concepts from questions, and 10,378 UMLS concepts from answers. Table 1 shows the total number of documents and sentences in our datasets (ie, blog postings, questions, answers). These numbers were compared to the “# with UMLS concepts,” the unique number of documents and sentences containing the identified UMLS concepts. Note that we can only extract concepts that are presented in UMLS. Thus, the total number of concepts in our datasets (which can include concepts that are not in UMLS) is not provided in Table 1.

There was a noticeable variation across the datasets. Over 80% of the documents from questions and answers contained 1 or more UMLS concepts whereas less than half of the documents from blogs did. Over half of the sentences from questions and answers contained at least 1 UMLS concept, while only 27% of those from blog posts contained at least 1 UMLS concept.

XSL•FO
RenderX

Table 1. Basic characteristics of UMLS concept coverage in our datasets.

	Tumblr		Yahoo! Answers			
	Blog postings		Questions		Answers	
	Total #	# with UMLS concepts	Total #	# with UMLS concepts	Total #	# with UMLS concepts
Document	3711	1388 (37.4%)	58,422	51,850 (88.8%)	58,422	51,550 (88.2%)
Sentence	47,413	12,802 (27.0%)	249,013	142,802 (57.3%)	348,793	216,736 (62.1%)
Concepts	–	2415	–	6452	–	10,378

Coverage by the UMLS Source Vocabularies

The concepts in the blogs were covered by 56 UMLS source vocabularies, while those in questions and answers were covered by 58 source vocabularies. Table 2 illustrates the top 20 most covered UMLS source vocabularies (The full names and the version information of the source vocabularies can be found in the Multimedia Appendix 1 Table-A1). SNOMED CT was dominant across all our datasets, ranging from 84.9% to 95.9%. It was followed by CHV (between 73.5% and 80.0%) and MTH (between 55.7% and 73.5%). Other source vocabularies within the top 10 for all of our datasets are: NCIt, Medical Subject Headings (MeSH), Computer Retrieval of Information on Scientific Projects Thesaurus (CSP), Library of Congress Subject Headings Northwestern University subset (LCH NW), Logical Observation Identifier Names and Codes (LOINC), and National Drug File – Reference Terminology (NDFRT), although the ranking order varies slightly across different datasets. Multimedia Appendix 1 Table-A2 provides example concepts in the top 3 most covered source vocabularies.

Table 2. Top 20 mostly covered UMLS source vocabularies.

	Tumblr			Yahoo! Answers					
Rank	Blogs (n=2415)			Questions (n=6452)			Answers (n=10,378)		
	Source vocabulary	# of concepts	%	Source vocabulary	# of concepts	%	Source vocabulary	# of concepts	%
1	SNOMED CT	2315	95.9	SNOMED CT	5476	84.9	SNOMED CT	9032	87.0
2	CHV	1931	80.0	CHV	4928	76.4	CHV	7625	73.5
3	MTH	1774	73.5	MTH	3899	60.4	MTH	5780	55.7
4	NCIt	1156	47.9	MeSH	2957	45.8	MeSH	4796	46.2
5	MeSH	1130	46.8	NCIt	2917	45.2	NCIt	4485	43.2
6	CSP	812	33.6	CSP	1840	28.5	NDFRT	2999	28.9
7	AOD	775	32.1	NDFRT	1775	27.5	CSP	2839	27.4
8	LCH_NW	771	31.9	LCH_NW	1627	25.2	LCH_NW	2436	23.5
9	LOINC	697	28.9	AOD	1585	24.6	AOD	2335	22.5
10	NDFRT	659	27.3	LOINC	1510	23.4	RXNORM	2099	20.2
11	LCH	587	24.3	RXNORM	1421	22.0	LOINC	2081	20.1
12	NCI_NCI-GLOSS	475	19.7	LCH	1187	18.4	LCH	1730	16.7
13	MEDLINEPLUS	402	16.6	NCI_NCI-GLOSS	952	14.8	NCI_FDA	1387	13.4
14	CST	365	15.1	NCI_FDA	868	13.5	DXP	1322	12.7
15	COSTAR	362	15.0	COSTAR	835	12.9	NCI_NCI-GLOSS	1321	12.7
16	NCI_FDA	345	14.3	DXP	830	12.9	COSTAR	1257	12.1
17	OMIM	342	14.2	CST	794	12.3	OMIM	1234	11.9
18	RXNORM	338	14.0	OMIM	790	12.2	CST	1206	11.6
19	DXP	326	13.5	MEDLINEPLUS	721	11.2	VANDF	1117	10.8
20	ICD9CM	241	10.0	VANDF	644	10.0	MTHSPL	1033	10.0

XSL•FO
RenderX

Table 3. Top 10 frequently observed concepts covered by both SNOMED CT and CHV.

Rank	Tumblr		Yahoo! Answers			
			Questions		Answers	
	Concept	Freq.	Concept	Freq.	Concept	Freq.
1	Blood (C0005767)	816	Blood (C0005767)	30,654	Blood (C0005767)	54,689
2	Pain (C0030193)	798	Sugars (C0242209)	29,593	Sugars (C0242209)	49,207
3	Insulin (C0021641)	744	Insulin (C0021641)	10,816	Insulin (C0021641)	27,887
4	Pharmaceutical preparations (C0013227)	719	Glucose (C0017725)	7394	Glucose (C0017725)	26,420
5	Sugars (C0242209)	699	Problem (C0033213)	5111	Pharmaceutical preparations (C0013227)	11,571
6	Disease (C0012634)	617	Water (C0043047)	4781	Diseases (C0012634)	9733
7	Problem (C0033213)	568	Pharmaceutical preparations (C0013227)	4456	Carbohydrates (C0007004)	9517
8	Diabetes mellitus (C0011849)	501	Hematologic tests (C0018941)	3784	Problem (C0033213)	9248
9	Teeth structure (C0040426)	424	Pain (C0030193)	3625	Water (C0043047)	5994
10	Operative surgery procedures (C0543467)	375	Urine (C0042036)	2550	Fasting (C0015663)	5848

Table 4. Top 10 frequently observed concepts covered by CHV but not SNOMED CT.

	Tumblr		Yahoo! Answers			
Rank			Questions		Answers	
	Concept (CUI)[a]	Freq.	Concept (CUI)	Freq.	Concept (CUI)	Freq.
1	Cider vinegar (C0937941)	54	Stomach (C0038351)	1050	Lantus (C0876064)	689
2	Apple cider vinegar (C1178459)	54	Lantus (C0876064)	571	Actos (C0875954)	659
3	Lantus (C0876064)	15	Humalog (C0528249)	260	Avandia (C0875967)	628
4	Gentle (C0720654)	11	NovoLog (C0939412)	180	HumaLog (C0528249)	289
5	Corrective (C0719519)	9	Glucophage (C0591573)	131	NovoLog (C0939412)	255
6	Botox (C0700702)	9	Levemir (C1314782)	122	Levemir (C1314782)	184
7	RID (C0073361)	6	Actos (C0875954)	95	Glucophage (C1314782)	161
8	HumaLog (C0528249)	5	Seroquel (C0287163)	78	Novolin (C0028467)	112
9	Bead Dosage Form (C0991566)	3	Synthroid (C0728762)	62	Viagra (C0663448)	105
10	Actos (C0875954)	3	Coumadin (C0699129)	54	Triphosphat (C0146894)	77

[a]CUI: concept unique identifier

There was significant overlap between the concepts from the top 2 source vocabularies, SNOMED CT and CHV 78.2% (1889/2415) concepts from blog postings, 70.0% (4518/6452) concepts in questions, and 68.4% (7095/10,378) concepts in answers. Table 3 shows the top 10 concepts. Note that we only show the preferred term of the concept in the UMLS throughout the paper. Diabetes-related concepts such as *Blood*, *Sugars*, *Insulin*, *Glucose*, and *Diabetes mellitus* were frequently mentioned (preferred names of a UMLS concept are denoted in italics). At the same time, it includes some general medical concepts such as *disease*, *pharmaceutical preparations*, and *problem*. Concepts related to glucose level in blood such as *blood*, *sugars*, *glucose*, and *carbohydrates* also appeared with high frequency.

A few concepts were only covered by CHV: 1.7% (40/2415) concepts in blog postings, 6.3% (409/6452) concepts in questions, and 5.1% (529/10,378) concepts in answers. Table 4 shows the top 10 most frequently observed UMLS concepts covered by CHV but not SNOMED CT in our datasets.

All the concepts in Table 4 are about pharmacological substances or organic chemicals, except *stomach* found within questions. Three concepts regarding insulin therapy for diabetes, such as *Lantus* (ie, insulin glargine injection), *Humalog* (ie, insulin lispro injection), and *Actos* (ie, pioglitazone hydrochloride) in blog postings and questions/answers appeared with high frequency. Diabetes-treatment-related concepts, such as *NovoLog* and *Glucophage*, are more frequently observed in questions and answers than blog postings. In total, 9 out of the

XSL•FO
RenderX

top 10 concepts in questions and answers were diabetes medications. Only 2 concepts, namely *stomach* in questions and *Viagra* in answers, are not directly related to diabetes treatment. On the other hand, some concepts in blogs were indirectly related to diabetes. For example, *cider vinegar*, *apple cider vinegar*, and *Botox* also frequently appeared.

There were also the concepts covered by SNOMED CT but not CHV: 17.6% (424/2415) concepts from blog postings, 957/6452 (14.8%) concepts in questions and 18.7% (1936/10,378) concepts in answers (See Table 5). Human body related concepts, such as *back structure excluding neck*, *entire heart*, *entire pancreas*, *entire kidney*, entire *skin*, and *entire eye*, were frequently used to describe their diabetes issues in blog postings or questions/answers. Three concepts, *entire skin*, *symptoms* and *fatty acid glycerol esters* were observed from all our datasets. *Massage* and *training* were frequently mentioned in blog postings, while *injection procedure* and *protective cup* were frequently mentioned in questions and answers but were not mentioned as frequently as in blog postings. As these concepts were frequently observed in social media, CHV should consider importing them to enrich its conceptual content.

Table 5. Top 10 frequently observed concepts covered by SNOMED CT but not CHV.

Rank	Tumblr		Yahoo! Answers			
			Questions		Answers	
	Concept (CUI)[a]	Freq.	Concept (CUI)	Freq.	Concept (CUI)	Freq.
1	Entire skin (C1278993)	524	Symptoms (C1457887)	7690	Symptoms (C1457887)	12,727
2	Symptoms (C1457887)	393	Fatty acid glycerol esters (C0015677)	1789	Fatty acid glycerol esters (C0015677)	8727
3	Back structure, excluding neck (C1995000)	236	Entire foot (C1281587)	1647	Entire cells (C1269647)	6435
4	Massage (C0024875)	217	Back structure, excluding neck (C1995000)	1589	Entire heart (C1281570)	3204
5	Fatty acid glycerol esters (C0015677)	210	Entire kidney (C1278978)	1368	Entire pancreas (C1278931)	3003
6	Training (C0220931)	163	Entire eye (C1280202)	1210	Entire skin (C1278993)	2614
7	Entire pancreas (C1278931)	157	Protective cup (C1533124)	1159	Protective cup (C1533124)	2178
8	Entire heart (C1281570)	156	Entire lower limb (C1269079)	985	Entire stomach (C1278920)	1876
9	Entire oral cavity (C1278910)	138	Entire hands (C1281583)	969	Injection procedure (C1533685)	1561
10	Entire spine (C1280065)	137	Entire skin (C1278993)	912	Entire bony skeleton (C1266909)	1501

[a]CUI: concept unique identifier

Semantic Types of the Identified Concepts

Among 127 UMLS semantic types (STY), about half of them were identified in our datasets: 52 STYs (40.9%) in the blog postings, 59 STYs (46.5%) in the questions, and 54 STYs (42.5%) in the answers. In general, there was a significant overlap of STYs across our datasets with 52 shared STYs. Seven STYs, however, were identified in the questions only, including “Functional Concept,” “Intellectual Product,” “Laboratory Procedure,” “Organ or Tissue Function,” “Organism Attribute,” “Social Behavior,” and “Substance.” Two STYs, “Fully Formed Anatomical Structure” and “Cell or Molecular Dysfunction,” were not found in questions, but in both the answer dataset and the blog dataset. Table 6 shows the top 20 frequent semantic types of the identified UMLS concepts in the different datasets respectively.

When comparing the top 10 frequently observed STYs across the datasets, 9 out of 10 STYs (ie, “Finding,” “Pharmacologic Substance,” “Therapeutic or Preventive Procedure,” “Disease or Syndrome,” “Organic Chemical,” “Body Part, Organ, or Organ Component,” “Sign or Symptom,” “Medical Device,” and “Amino Acid, Peptide, or Protein”) commonly appeared across the datasets with minor differences in terms of frequency. “Laboratory Procedure” appeared frequently in questions but not in blogs and answers. “Pathologic Function” appeared frequently in answers but not in blogs and questions. Example concepts of the frequently observed STYs showed that laypeople tend to frequently use common concepts to describe their diabetes-related issues in social media. To illustrate, *Sugars*, *Insulin*, *Glucose* ranked in top 5 concepts of the STY “Pharmacologic Substance.” Similarly, the concepts such as *Disease* and *Communicable Diseases* appeared frequently among the concepts of the STY “Disease or Syndrome.” We provide the top 5 frequent concepts for the top 10 frequently observed semantic types in Multimedia Appendix 1 Table A3.

XSL•FO
RenderX

Table 6. Top 20 frequently observed semantic types of the identified concepts.

Rank	Tumblr			Yahoo! Answers					
	Blogs			Questions			Answers		
	Semantic type	Concept[a]		Semantic type	Concept		Semantic type	Concept	
		n (%)	Freq.		n (%)	Freq.		n (%)	Freq.
1	Finding	380 (15.7)	5277	Pharmacologic substance	1240 (19.2)	53,976	Pharmacologic substance	1995 (19.2)	185,880
2	Pharmacologic substance	307 (12.7)	4413	Organic chemical	1006 (15.6)	41,255	Organic chemical	1692 (16.3)	123,509
3	Therapeutic or preventive procedure	241 (10.0)	3184	Finding	895 (13.9)	30,458	Disease or syndrome	1511 (14.6)	57,379
4	Disease or syndrome	239 (9.9)	2923	Disease or syndrome	743 (11.5)	28,041	Finding	1302 (12.5)	76,765
5	Organic chemical	225 (9.3)	2737	Body part, organ, or organ component	484 (7.5)	27,172	Body part, organ, or organ component	666 (6.4)	48,584
6	Body part, organ, or organ component	208 (8.6)	2566	Sign or symptom	338 (5.2)	19,601	Therapeutic or preventive procedure	583 (5.6)	16,555
7	Sign or symptom	145 (6.0)	2214	Therapeutic or preventive procedure	331 (5.1)	16,372	Amino acid, peptide, or protein	495 (4.8)	40,521
8	Medical device	134 (5.5)	1319	Amino acid, peptide, or protein	305 (4.7)	13,178	Sign or symptom	436 (4.2)	38,905
9	Amino acid, peptide, or protein	70 (2.9)	1112	Medical device	201 (3.1)	12,862	Medical device	347 (3.3)	20,391
10	Biologically active substance	69 (2.9)	1093	Laboratory procedure	180 (2.8)	10,580	Pathologic function	292 (2.8)	12,551

[a] The percentage was calculated based on the total number of unique identified UMLS concepts: blogs in Tumblr: n=2415, questions in Yahoo! Answers: n=6452, answers in Yahoo! Answers: n=10,378

Discussion

Principal Findings

Previous studies [12-15] utilized user-generated documents including social media. However, they mainly used a single test bed based on the assumption that the selected test bed would properly reflect people's medical concepts. Our study involved different types of social media which contains texts that laypeople generated for different purposes: questions are for expressing their health information seeking needs; blogs and answers are more likely for sharing their knowledge, experiences, and opinions to others. The current study investigated the terminology coverage in consumer-generated text in social media by identifying UMLS concepts and their semantic types. Our findings demonstrated that consumers use medical concepts not only from controlled vocabularies developed for consumers (ie, CHV) but also those for health professionals (eg, SNOMED CT). Our results are in line with prior observations that consumers use both lay and professional terms [24,26,65] and demonstrated that CHV can be enriched by other source vocabularies in the UMLS.

The UMLS concept usage in blogs and social Q&A was different in that the UMLS concepts appeared more frequently in the postings of social Q&A (almost 90% questions and answers) in comparison to blog postings (about 30%). Social Q&A users mainly discuss health-related issues (in the current study, diabetes-related issues) in their postings, because their participation in question asking and answering is purpose-driven. On the other hand, blog users often elaborate nonhealth related topics in their postings, although they tagged their postings with "diabetes."

In spite of the differences of the overall UMLS concept coverage between blogs and social Q&A, we found that the UMLS concepts identified in different datasets can be covered by a similar number of UMLS source vocabularies. Two UMLS source vocabularies, ie, SNOMED CT and CHV, showed the best coverage. Social media users in our datasets may have advanced medical knowledge because they often use professional terms. CHV demonstrated the second largest coverage for all the datasets despite the fact that CHV has a much smaller number of concepts and terms than SNOMED CT (1:6 ratio). CHV was developed to incorporate consumer expressions presented in consumer-generated text data. Our findings showed that different social media platforms may play a similar role as consumer-generated documents for CHV enrichment, which confirmed the literature [66].

A comparison of the concept coverage between SNOMED CT and CHV in our datasets led us to examine the difference between the concept usages among blog and social media users. For example, *cider vinegar*, *apple cider vinegar*, *massages*, and *training* were frequently mentioned in blog postings, while they were not frequently mentioned in questions and answers. However, concepts pertaining to insulin therapy, such as *Lantus*, *Humalog*, and *Actos*, are frequently used in questions/answers. Consumers often inquire about a variety of insulin therapies in social Q&A, while blogs often include recipes specific to the use of *vinegar*, a popular ingredient in diabetes-controlling food. *Botox* and *Viagra* were often mentioned in blog postings and answers. They could be important for diabetic patients, although they may not be closely related to control diabetes. It would be interesting to further analyze the relationship of these terms to diabetes. An in-depth analysis of the identified concepts along with how they are used in the original postings could produce useful information for understanding consumers' information needs and use.

According to our analysis, the percentage of unique concepts covered by CHV but not by SNOMED CT varied from 1.7% to 6.3%. In the blog dataset, where approximately 3000 blogs were analyzed, only 40 concepts were covered by CHV exclusively. On the other hand, in Yahoo! Answers, 409 concepts (6.3%) in questions and 529 concepts (5.1%) in answers were covered by CHV but not by SNOMED CT. These results indicate that the larger datasets would yield more lay concepts. The size of dataset also appeared to affect the diversity of semantics. The same set of 9 semantic types was observed frequently in all our datasets. "Finding," "Pharmacologic Substance," and "Disease and Syndrome" were among the top 4 most observed semantic types.

Differences were observed as well. Blogs might be better platforms for consumers to discuss organic chemical, pharmacologic substances, or therapeutic or preventive procedure for diabetes. Yet, concepts of organic chemical and pharmacologic substances were also frequently used in social Q&A. In social Q&A data, 7 semantic types that were not identified in blogs were observed, indicating that larger datasets may yield more diverse medical concepts.

Limitations

This study has a few limitations. First, the blog data in Tumblr and Yahoo! Answers data were collected in different time frames and are different in size, which might have affected the findings of this study. Smaller volumes of blog data used in this study may affect the diversity of the UMLS concepts identified. Even though blogging and question posting/answering are dynamic online activities for those living with chronic diseases, Tumblr and Yahoo! Answers may not represent all the health information users' concept usage. The datasets could be expanded to include other types of social media such as diabetes-related discussion boards. The users of these online sources may be biased towards those with greater technological proficiency, such as those who are younger, more educated or those in a higher socioeconomic status who are more likely to seek health information on the Internet. This study may not reflect the experiences of those who are older adults, less educated or underprivileged [67]. Second, even though the automated NLP techniques that were employed in the current study were cost-effective, direct interaction with ordinary health information users would allow the researchers to capture more accurate meaning of medical concepts that these individuals commonly use to describe their health issues. Moreover, a qualitative approach such as content analysis also would help to identify contextual semantics of the concepts. Third, although the WSD function of YTEX is effective in selecting the most possible UMLS concepts for a term in free text, the same term may be matched to different ambiguous UMLS concepts. This is mainly due to the fact that the UMLS may contain unmapped synonymous concepts. Ideally, manual review by domain experts could be applied to further refine the automatic mapping results.

Conclusions

The current study examined the potential of social media as user-generated documents in which consumers' medical concepts can be observed and leveraged for controlled vocabulary development for ordinary health information users. We selected and tested 2 social media venues, namely blogs and social Q&A. Our findings showed stronger similarities rather than differences in the controlled vocabulary usage. The size of a dataset may affect the number of concepts identified. However, the similarities in the source vocabularies, frequently used concepts, and semantic types of the concepts indicate that social media sites tend to reflect the common sense of laypeople. More importantly, we found that social media users not only employ consumer concepts in CHV but also concepts in professional vocabularies such as SNOMED CT. This indicates that CHV still has room for improvements by incorporating concepts from other UMLS source vocabularies. The focus of our study is not to identify a list of consumer medical concepts, but to test the feasibility of leveraging social media data to identify consumer concepts covered by existing UMLS source vocabularies. Ultimately, it would assist consumers' health information searches online, narrowing the disparity between ordinary health information users and medical professionals. In future studies, we will employ automated approaches to identify and recommend new medical terms and concepts from social media to enrich CHV.

Acknowledgments

We would like to thank Dr Warren Allen for providing the computing resources for this work. This work was partially supported by an Amazon Web Services Education and Research Grant Award (PI: He). The work was also partially supported by National Center for Advancing Translational Sciences under the Clinical and Translational Science Award UL1TR001427 (PI: Nelson & Shenkman). The content is solely the responsibility of the authors and does not represent the official view of the National Institutes of Health.

XSL•FO
RenderX

Authors' Contributions

MP initiated the idea of this study. ZH led the conceptualization, design, and implementation of this study. MP collected and provided the blog data from Tumblr.com. SO collected and provided the social Q&A data from Yahoo! Answers. ZC performed the natural language processing on the datasets and structured the results in a relational database. MP performed the data analysis and drafted the initial version; ZH, SO, BJ extensively revised the draft critically and iteratively for important intellectual content. All authors contributed to the methodology development, results interpretation, edited the paper significantly, and gave final approval for the version to be published. ZH takes primary responsibility for the research reported here.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Table A1. Full names of the UMLS source vocabularies in Table 2. Table A2. Top 10 frequently observed concepts in the top 3 most covered source vocabularies. Table A3. Top 5 most frequently concepts in the top 9 frequent semantic types.

[PDF File (Adobe PDF File), 125KB - medinform_v4i4e41_app1.pdf]

References

1. Messai R, Simonet M, Bricon-Souf N, Mousseau M. Characterizing consumer health terminology in the breast cancer field. Stud Health Technol Inform 2010;160(Pt 2):991-994. [Medline: 20841832]
2. Poikonen T, Vakkari P. Lay persons? and professionals? nutrition-related vocabularies and their matching to a general and a specific thesaurus. Journal of Information Science 2009;35(2):232-243.
3. Smith CA, Wicks PJ. PatientsLikeMe: Consumer health vocabulary as a folksonomy. AMIA Annu Symp Proc 2008:682-686 [FREE Full text] [Medline: 18999004]
4. Patrick TB, Monga HK, Sievert ME, Houston HJ, Longo DR. Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes. J Med Internet Res 2001;3(3):E24 [FREE Full text] [doi: 10.2196/jmir.3.3.e24] [Medline: 11720966]
5. Zielstorff RD. Controlled vocabularies for consumer health. J Biomed Inform 2003;36(4-5):326-333 [FREE Full text] [Medline: 14643728]
6. Tse T, Soergel D. Exploring medical expressions used by consumers and the media: an emerging view of consumer health vocabularies. AMIA Annu Symp Proc 2003:674-678 [FREE Full text] [Medline: 14728258]
7. Vydiswaran VG, Mei Q, Hanauer DA, Zheng K. Mining consumer health vocabulary from community-generated text. AMIA Annu Symp Proc 2014;2014:1150-1159 [FREE Full text] [Medline: 25954426]
8. Seedorff M, Peterson KJ, Nelsen LA, Cocos C, McCormick JB, Chute CG, et al. Incorporating expert terminology and disease risk factors into consumer health vocabularies. Pac Symp Biocomput 2013:421-432 [FREE Full text] [Medline: 23424146]
9. Gross T, Taylor A. What have we got to lose? The effect of controlled vocabulary on keyword searching results. College & Research Libraries 2005;66(3):212-230.
10. Lewis D, Eysenbach G, Jimison H, Kukafka R, Stavri P. Consumer health informatics. In: Consumer Health Informatics: Informing Consumers and Improving Health Care. New York, NY: Springer; 2005:1-7.
11. Roberts K, Demner-Fushman D. Interactive use of online health resources: a comparison of consumer and professional questions. J Am Med Inform Assoc 2016 Jul;23(4):802-811. [doi: 10.1093/jamia/ocw024] [Medline: 27147494]
12. Zeng Q, Kogan S, Ash N, Greenes RA. Patient and clinician vocabulary: how different are they? Stud Health Technol Inform 2001;84(Pt 1):399-403. [Medline: 11604772]
13. Plovnick RM, Zeng QT. Reformulation of consumer health queries with professional terminology: a pilot study. J Med Internet Res 2004 Sep 03;6(3):e27 [FREE Full text] [doi: 10.2196/jmir.6.3.e27] [Medline: 15471753]
14. Brennan PF, Aronson AR. Towards linking patients and clinical information: detecting UMLS concepts in e-mail. J Biomed Inform 2003;36(4-5):334-341. [Medline: 14643729]
15. Smith CA, Stavri PZ, Chapman WW. In their own words? A terminological analysis of e-mail to a cancer information service. Proc AMIA Symp 2002:697-701 [FREE Full text] [Medline: 12463914]
16. Harpring P. What Are Controlled Vocabularies? In: Baca M, editor. Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works. Los Angeles, CA: Getty Publications; 2010.
17. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc 2006;13(1):24-29 [FREE Full text] [doi: 10.1197/jamia.M1761] [Medline: 16221948]
18. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998 Nov;37(4-5):394-403 [FREE Full text] [Medline: 9865037]
19. Arts DG, Cornet R, de Jonge E, de Keizer NF. Methods for evaluation of medical terminological systems--a literature review and a case study. Methods Inf Med 2005;44(5):616-625. [Medline: 16400369]

XSL•FO
RenderX

20. Greenberg J. Metadata and the World Wide Web. In: Drake M, editor. Encyclopedia of Library and Information Science. New York, NY: Marcel Deker, Inc; 2003:1876-1888.
21. Mathes A. Folksonomies - cooperative classification and communication through shared metadata. 2004 Dec. URL: http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html [accessed 2016-11-02] [WebCite Cache ID 6lilnyrVJ]
22. Kim S. An exploratory study of user-centered indexing of published biomedical images. J Med Libr Assoc 2013 Jan;101(1):73-76 [FREE Full text] [doi: 10.3163/1536-5050.101.1.011] [Medline: 23405049]
23. MacLean DL, Heer J. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. J Am Med Inform Assoc 2013;20(6):1120-1127 [FREE Full text] [doi: 10.1136/amiajnl-2012-001110] [Medline: 23645553]
24. 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 May 17;13(2):e37 [FREE Full text] [doi: 10.2196/jmir.1636] [Medline: 21586386]
25. Hicks A, Hogan WR, Rutherford M, Malin B, Xie M, Fellbaum C, et al. Mining Twitter as a First Step toward Assessing the Adequacy of Gender Identification Terms on Intake Forms. AMIA Annu Symp Proc 2015;2015:611-620 [FREE Full text] [Medline: 26958196]
26. Lewis D, Brennan PF, McCray AT, Tuttle M, Bachman J. If We Build It, They Will Cometandardized Consumer Vocabularies. Studies in Health Technology and Informatics 2001;84:1530 [FREE Full text]
27. Kamdar MR, Tudorache T, Musen M. A Systematic Analysis of Term Reuse and Term Overlap across Biomedical Ontologies. Semantic Web – Interoperability, Usability, Applicability 2016 (forthcoming). [FREE Full text]
28. He Z, Geller J, Chen Y. A comparative analysis of the density of the SNOMED CT conceptual content for semantic harmonization. Artif Intell Med 2015 May;64(1):29-40 [FREE Full text] [doi: 10.1016/j.artmed.2015.03.002] [Medline: 25890688]
29. He Z, Geller J, Elhanan G. Categorizing the Relationships between Structurally Congruent Concepts from Pairs of Terminologies for Semantic Harmonization. AMIA Jt Summits Transl Sci Proc 2014;2014:48-53 [FREE Full text] [Medline: 25717400]
30. He Z, Chen Y, de Coronado S, Piskorski K, Geller J. Topological-Pattern-based Recommendation of UMLS Concepts for National Cancer Institute Thesaurus. AMIA Annu Symp Proc 2016 (forthcoming).
31. Chandar P, Yaman A, Hoxha J, He Z, Weng C. Similarity-Based Recommendation of New Concepts to a Terminology. AMIA Annu Symp Proc 2015;2015:386-395 [FREE Full text] [Medline: 26958170]
32. 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]
33. World Health Organization. Global Report on Diabetes. World Health Organization. URL: http://apps.who.int/iris/bitstream/10665/204871/1/9789241565257_eng.pdf?ua=1 [accessed 2016-11-02] [WebCite Cache ID 6limIzVBr]
34. American Diabetes Association. Diabetes complications. URL: http://www.diabetes.org/living-with-diabetes/complications/ [accessed 2016-11-02] [WebCite Cache ID 6limMPM2T]
35. Krug EG. Trends in diabetes: sounding the alarm. Lancet 2016 Apr 9;387(10027):1485-1486 [FREE Full text] [doi: 10.1016/S0140-6736(16)30163-5] [Medline: 27061675]
36. Fox S. The social life of health information. Pew Research Center. URL: http://www.pewresearch.org/fact-tank/2014/01/15/the-social-life-of-health-information/ [accessed 2016-11-02] [WebCite Cache ID 6limSyxAR]
37. Andersen NB, Söderqvist T. Social media and public health research. Copenhagen, Denmark: University of Copenhagen. 2012 Aug 20. URL: http://www.museion.ku.dk/wp-content/uploads/FINAL-Social-Media-and-Public-Health-Research.pdf [accessed 2016-11-02] [WebCite Cache ID 6limaWFsM]
38. Oh S. The characteristics and motivations of health answerers for sharing information, knowledge, and experiences in online environments. J. Am. Soc. Inf. Sci 2011 Nov 01;63(3):543-557. [doi: 10.1002/asi.21676]
39. Giustini D. How Web 2.0 is changing medicine. BMJ 2006 Dec 23;333(7582):1283-1284 [FREE Full text] [doi: 10.1136/bmj.39062.555405.80] [Medline: 17185707]
40. U.S. National Library of Medicine. Fact Sheet of the UMLS Metathesaurus. URL: https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html [accessed 2016-11-02] [WebCite Cache ID 6limjdjWl]
41. U.S. National Library of Medicine. SNOMED Clinical Terms. URL: https://www.nlm.nih.gov/healthit/snomedct/index.html [accessed 2016-11-02] [WebCite Cache ID 6limqaGNJ]
42. Finnegan R. ICD-9-CM coding for physician billing. J Am Med Rec Assoc 1989 Feb;60(2):22-23. [Medline: 10303229]
43. Bennett CC. Utilizing RxNorm to support practical computing applications: capturing medication history in live electronic health records. J Biomed Inform 2012 Aug;45(4):634-641 [FREE Full text] [doi: 10.1016/j.jbi.2012.02.011] [Medline: 22426081]
44. U.S. National Library of Medicine. Fact Sheet of the UMLS Semantic Network. URL: https://www.nlm.nih.gov/pubs/factsheets/umlssemn.html [accessed 2016-07-13] [WebCite Cache ID 6iyTFIvxE]
45. Zeng-Treitler Q, Goryachev S, Kim H, Keselman A, Rosendale D. Making texts in electronic health records comprehensible to consumers: a prototype translator. AMIA Annu Symp Proc 2007:846-850 [FREE Full text] [Medline: 18693956]
46. Zielstorff R. Controlled vocabularies for consumer health. Journal of Biomedical Informatics 2003 Aug;36(4-5):326-333. [doi: 10.1016/j.jbi.2003.09.015]

XSL•FO
RenderX

47. Miller EA, Pole A. Diagnosis blog: checking up on health blogs in the blogosphere. Am J Public Health 2010 Aug;100(8):1514-1519. [doi: 10.2105/AJPH.2009.175125] [Medline: 20558802]
48. Lagu T, Kaufman EJ, Asch DA, Armstrong K. Content of weblogs written by health professionals. J Gen Intern Med 2008 Oct;23(10):1642-1646 [FREE Full text] [doi: 10.1007/s11606-008-0726-6] [Medline: 18649110]
49. Boulos Maged N Kamel, Maramba I, Wheeler S. Wikis, blogs and podcasts: a new generation of Web-based tools for virtual collaborative clinical practice and education. BMC Med Educ 2006;6:41 [FREE Full text] [doi: 10.1186/1472-6920-6-41] [Medline: 16911779]
50. Oomen-Early J, Burke S. Entering the Blogosphere: Blogs as Teaching and Learning Tools in Health Education. International Electronic Journal of Health Education 2007;10:186-196 [FREE Full text]
51. Cobus L. Using blogs and wikis in a graduate public health course. Med Ref Serv Q 2009;28(1):22-32. [doi: 10.1080/02763860802615922] [Medline: 19197741]
52. Shah C, Oh S, Oh J. Research agenda for social Q&A. Library & Information Science Research 2009;31(4):205-209. [doi: 10.1016/j.lisr.2009.07.006]
53. Statista. Statistics and facts about Tumblr. URL: http://www.statista.com/topics/2463/tumblr/ [accessed 2016-02-06] [WebCite Cache ID 6iyTGqcPl]
54. DearBlogger: The Blogging Answers Community. The Best Places to Start a Blog. URL: http://www.dearblogger.org/blogger-or-wordpress-better [accessed 2016-11-02] [WebCite Cache ID 6linVHtiQ]
55. Fitzpatrick J. Lifehacker. 2010 Jun 20. Five Best Blogging Platforms. Lifehacker. 2010 Jun 20. URL: http://lifehacker.com/5568092/five-best-blogging-platforms [accessed 2016-11-02] [WebCite Cache ID 6linqjSUQ]
56. Quantcast. Statistics of Yahoo! Answers. URL: https://www.quantcast.com/answers.yahoo.com [accessed 2016-07-13] [WebCite Cache ID 6iyTJNzkf]
57. Temmerman R. Towards new ways of terminology description: The sociocognitive-approach. In: Towards new ways of terminology description: the sociocognitive-approach. Amsterdam: J. Benjamins; 2000.
58. Crystal D. Dictionary of Linguistics and Phonetics (The Language Library). Hoboken, NJ: Wiley-Blackwell; Jun 2008.
59. U.S. National Library of Medicine. UMLS Glossary. URL: https://www.nlm.nih.gov/research/umls/new_users/glossary.html [accessed 2016-01-08] [WebCite Cache ID 6iyTK76nW]
60. Apache Software Foundation. cTAKES (clinical Text Analysis and Knowledge Extraction System). 2016 Jan 18. URL: http://ctakes.apache.org [accessed 2016-01-18] [WebCite Cache ID 6iyTLd7aW]
61. Garla V, Lo RV, Dorey-Stein Z, Kidwai F, Scotch M, Womack J, et al. The Yale cTAKES extensions for document classification: architecture and application. J Am Med Inform Assoc 2011;18(5):614-620 [FREE Full text] [doi: 10.1136/amiajnl-2011-000093] [Medline: 21622934]
62. Baldridge J. Apache Software Foundation. The openNLP project. URL: http://opennlp.apache.org/index [accessed 2016-11-02] [WebCite Cache ID 6lioR15Gv]
63. Apache Software Foundation. The OpenNLP Documentation. URL: https://opennlp.apache.org/documentation.html [accessed 2016-11-02] [WebCite Cache ID 6lioZxO8m]
64. Garla VN, Brandt C. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC Bioinformatics 2012 Oct 10;13:261 [FREE Full text] [doi: 10.1186/1471-2105-13-261] [Medline: 23046094]
65. He Z, Park M, Chen Z. UMLS-Based Analysis of Medical Terminology Coverage for Tags in Diabetes-Related Blogs. In: Philadelphia, PA. 2016 Presented at: iConference 2016; March 20-23, 2016 URL: https://www.ideals.illinois.edu/handle/2142/89441
66. Sarasohn-Kahn J. The Wisdom of Patients: Health Care Meets Online Social Media. California Health Care Foundation. 2008 Apr. URL: http://www.chcf.org/publications/2008/04/the-wisdom-of-patients-health-care-meets-online-social-media [accessed 2016-11-02] [WebCite Cache ID 6liosAzP6]
67. Zhang Y. Beyond quality and accessibility: Source selection in consumer health information searching. J Assn Inf Sci Tec 2014 Jan 07;65(5):911-927. [doi: 10.1002/asi.23023]

Abbreviations

APIs: Application Program Languages
AUI: atom unique identifier
CSP: Computer Retrieval of Information on Scientific Projects Thesaurus
CUI: concept unique identifier
IC: information content
LCH NW: Library of Congress Subject Headings, Northwestern University subset
LOINC: Logical Observation Identifier Names and Codes
MeSH: Medical Subject Headings
NCIt: National Cancer Institute Thesaurus
NDFRT: National Drug File – Reference Terminology
NER: named entity recognition

XSL•FO
RenderX

NLM: National Library of Medicine
NLP: natural language processing
OAC CHV: Open-Access Collaborative Consumer Health Vocabulary
POS: Part-Of-Speech
Q&A: questions and answers
SNOMED CT: SNOMED Clinical Terms
STY: semantic type
UIMA: Unstructured Information Management Architecture
UMLS: Unified Medical Language System
WSD: word sense disambiguation

Edited by G Eysenbach; submitted 10.03.16; peer-reviewed by C Kruse, H Wu, S Kim; comments to author 28.06.16; revised version received 02.08.16; accepted 22.10.16; published 24.11.16.

Please cite as:
Park MS, He Z, Chen Z, Oh S, Bian J
Consumers' Use of UMLS Concepts on Social Media: Diabetes-Related Textual Data Analysis in Blog and Social Q&A Sites
JMIR Med Inform 2016;4(4):e41
URL: https://medinform.jmir.org/2016/4/e41/
doi:10.2196/medinform.5748
PMID:27884812

©Min Sook Park, Zhe He, Zhiwei Chen, Sanghee Oh, Jiang Bian. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 24.11.2016. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Population Analysis of Adverse Events in Different Age Groups Using Big Clinical Trials Data

Jake Luo[1,2,3], PhD; Christina Eldredge[2,4], MD; Chi C Cho[3,5], MS; Ron A Cisler[1,2,3,6,7,8], PhD

[1]Center for Biomedical Data and Language Processing, College of Health Sciences, Department of Health Informatics and Administration, University of Wisconsin-Milwaukee, Milwaukee, WI, United States

[2]Department of Health Informatics and Administration, College of Health Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI, United States

[3]College of Health Science, University of Wisconsin-Milwaukee, Milwaukee, WI, United States

[4]Department of Family and Community Medicine, Medical College of Wisconsin, Milwaukee, WI, United States

[5]Center for Aging and Translational Research, University of Wisconsin-Milwaukee, Milwaukee, WI, United States

[6]Center for Urban Population Health, Milwaukee, WI, United States

[7]School of Medicine and Public Health, University of Wisconsin-Madison, Milwaukee, WI, United States

[8]Zilber School of Public Health, University Wisconsin-Milwaukee, Milwaukee, WI, United States

Corresponding Author:
Jake Luo, PhD
Center for Biomedical Data and Language Processing
College of Health Sciences, Department of Health Informatics and Administration
University of Wisconsin-Milwaukee
2025 E Newport Avenue, NWQ-B Room 6469
Milwaukee, WI,
United States
Phone: 1 414 229 7333
Fax: 1 414 229 2619
Email: jakeluo@uwm.edu

Abstract

Background: Understanding adverse event patterns in clinical studies across populations is important for patient safety and protection in clinical trials as well as for developing appropriate drug therapies, procedures, and treatment plans.

Objectives: The objective of our study was to conduct a data-driven population-based analysis to estimate the incidence, diversity, and association patterns of adverse events by age of the clinical trials patients and participants.

Methods: Two aspects of adverse event patterns were measured: (1) the adverse event incidence rate in each of the patient age groups and (2) the diversity of adverse events defined as distinct types of adverse events categorized by organ system. Statistical analysis was done on the summarized clinical trial data. The incident rate and diversity level in each of the age groups were compared with the lowest group (reference group) using *t* tests. Cohort data was obtained from ClinicalTrials.gov, and 186,339 clinical studies were analyzed; data were extracted from the 17,853 clinical trials that reported clinical outcomes. The total number of clinical trial participants was 6,808,619, and total number of participants affected by adverse events in these trials was 1,840,432. The trial participants were divided into eight different age groups to support cross-age group comparison.

Results: In general, children and older patients are more susceptible to adverse events in clinical trial studies. Using the lowest incidence age group as the reference group (20-29 years), the incidence rate of the 0-9 years-old group was 31.41%, approximately 1.51 times higher (*P*=.04) than the young adult group (20-29 years) at 20.76%. The second-highest group is the 50-59 years-old group with an incidence rate of 30.09%, significantly higher (*P*<.001) when compared with the lowest incidence in the 20-29 years-old group. The adverse event diversity also increased with increase in patient age. Clinical studies that recruited older patients (older than 40 years) were more likely to observe a diverse range of adverse events (*P*<.001). Adverse event diversity increased at an average rate of 77% for each age group (older than 30 years) until reaching the 60-69 years-old group, which had a diversity level of 54.7 different types of adverse events per trial arm. The 70-100 years-old group showed the highest diversity level of 55.5 events per trial arm, which is approximately 3.44 times more than the 20-29 years-old group (*P*<.001). We also observe that adverse events display strong age-related patterns among different categories.

XSL•FO
RenderX

Conclusion: The results show that there is a significant adverse event variance at the population level between different age groups in clinical trials. The data suggest that age-associated adverse events should be considered in planning, monitoring, and regulating clinical trials.

(JMIR Med Inform 2016;4(4):e30) doi:10.2196/medinform.6437

KEYWORDS

big data analysis; adverse events; clinical trial data; population health; clinical trial safety; data processing and integration

Introduction

Clinical trials explore and evaluate the safety and effectiveness of clinical interventions. Many clinical trial interventions are experimental, and thus they have greater risks to adversely affect the health of the participants in comparison to standard clinical practice. The adverse events data in this study were extracted from ClinicalTrials.gov. An adverse event is defined by ClincialTrials.gov as unfavorable changes in health during clinical trials, including abnormal laboratory findings [1]. Serious adverse events include events that result in death, disability, birth defects, inpatient hospitalizations, prolongation of hospitalization, or life-threatening conditions.

Adverse event reporting is a critical measurement for estimating the safety of new treatments and new drug therapies. According to the literature, adverse events could be one of the leading causes of death in the United States [2]. Serious adverse events could lead to hospitalization, life-threatening symptoms, or even patient death [3,4]. Studies also found that adverse reactions are a significant cause of injury in children [4]. Therefore, analyzing the pattern of adverse events in clinical studies has a great importance to public health and significant value to clinical research.

Drug studies have shown the importance of age as a factor influencing the incidence of adverse events in clinical studies. For example, among heart failure patients, the adverse event incidence of digoxin increases progressively with age, from 1.7% among patients younger than 50 years old to 5.4% among patients aged older than 80 years. Hospitalizations from digoxin toxicity also increase with age [5]. Additionally, a recent study found that age and obesity are significant risk factors for adverse events after hip arthroplasty treatment [6]. Furthermore, a clinical trial involving inhaled corticosteroids for treating children with asthma found that cough and perioral dermatitis are more frequent in children younger than 6 years old, while hoarseness is more frequent in older children [7]. These studies analyzed the association between individual treatments and adverse events. However, currently there is a lack of population health level analysis of adverse event association with patient age.

The objective of this study was to compare the incidence rate and diversity of adverse events in clinical trials among different age groups, revealing potential adverse event disparities between the patient age groups. In comparison to standard adverse event analyses in individual clinical trials, this study focused on comparing adverse events between different age groups across 17,853 trials and 6,808,619 participants that could have different interventions. The adverse events observed during clinical trials are not necessarily induced by the trial intervention. The adverse events could be inherited from the recruited participant population, which is a crucial factor to consider for planning and conducting clinical trials. We aimed to reveal adverse event risks and patterns on the population level across different age groups in clinical trials to inform investigators for use in future clinical trial preparation. Currently, there is a gap in this level of knowledge.

Methods

Data Extraction and Preparation

The source data in this study were extracted from ClinicalTrials.gov, which is the largest public clinical trial repository [1,8]. We downloaded 186,339 clinical trial studies submitted from 2000 to 2014, from which we extracted 17,853 studies that published the actual outcome results. Using an XML parser [9] developed to extract data elements from clinical trial reports, we collected the clinical trial title, sponsor type, intervention, participant age, arm group, and adverse events. In this study, we focused on analyzing the age categories and their association with adverse events.

We collected a total of 6,808,619 clinical trial participants and approximately 11,000 types of adverse events. Based on the reported mean age, the study population in each of the trial arms was categorized into eight age groups in 10-year increments except for the 70 to 100 years-old group. The adverse events were originally encoded in different terminologies in the reports, such as the Medical Dictionary for Regulatory Activities (MedDRA) [10,11], the World Health Organization Adverse Reactions Terminology [12], and the *International Classification of Diseases, Ninth Revision* [13]. We mapped the extracted reported adverse events into Unified Medical Language System (UMLS)-based standardized concepts using the MetaMap application [14,15] to normalize the terminology across different trials. All collected adverse events were classified into the 26-group MedDRA system organ classes (SOCs) [10] based on the classification provided by ClinicalTrials.gov. The extracted data were stored in the Clinical Trial Adverse Event Database for analysis [9]. We analyzed the association between age and adverse event from two perspectives: the incidence rate of adverse events and the diversity of adverse events. Statistical analysis was done on the summarized data using the Excel 2016 (Microsoft Corp) statistical package.

Analysis of Adverse Event Incidence

The incidence of adverse events is commonly used to evaluate the safety of a new treatment. If an adverse event has a high incidence rate in clinical trials, this indicates the event is more likely to occur among the study patients. For each of the adverse

XSL•FO
RenderX

events in each trial arm, we collected the total number of affected patients and the number of at-risk patients. The incidence per study is calculated as the percentage of at-risk patients affected by adverse events. Many population-level studies have analyzed the adverse event incidence rate among different age groups and clinical settings such as in-hospital, outpatient, after discharge, and nursing homes. However, in the past, research on adverse events in clinical trial studies has primarily focused on individual drug and selected participant groups. Population-level analysis of adverse events across different age groups and interventions in clinical trials is lacking. The objective of this study is to fill this knowledge gap by providing systematic analysis of adverse event incidence in clinical trials by comparing the incidence and diversity of adverse events in different age groups across clinical trials.

Analysis of Adverse Event Diversity

Adverse event diversity examines how many distinct types of adverse events (eg, cardiac failure, depression, patient death) occur in clinical studies. The diversity of adverse event occurrence is an important factor for estimating intervention risks; however, it is often overlooked. When a study therapy is associated with a high diversity of adverse events in a population group, the complexity and cost of developing effective procedures to prevent and treat the adverse events could also increase [16,17]. To compare the adverse event diversity, we categorized the participant population in each of the trial arms according to the age groups. Then, we summarized the distinct types of adverse events that occurred in each of the age groups in the trial. The adverse event diversity was calculated on the trial arm level. For example, if a trial arm for a study has the adverse events heart failure, dizziness, and nausea, then the diversity of this trial arm would be 3. The mean of the adverse event diversity in each of the age groups was calculated as the number of distinct adverse event types divided by the number of trial arms of the age group, which indicates the average number of distinct adverse events in each of the trial arms. To further assess adverse event diversities in different organ systems, the adverse events were categorized into the 26 MedDRA SOCs. We then compared the adverse event diversity in each of the organ classes across the eight age groups.

Results

Incidence of Adverse Event

Figure 1 shows the average adverse event incidence rate of each of the age groups. The total number of affected patients and the corresponding MedDRA SOCs are also shown in Figure 1. The results show that the 20 to 29 years-old group has the lowest adverse event incidence rate of 20.76%. The highest group is the 0 to 9 years-old group, with an incidence rate of 31.41% and *P*=.02 (*P*<.05, *t* test) when compared with the lowest group, 20 to 29 years-old. The risk difference between the 0 to 9 years-old and 20 to 29 years-old groups is 10.6% (SE 0.00070). The results indicate that young children are more susceptible to adverse events than the young adult reference groups on a population level. The second highest group is 50 to 59 years-old, with an incidence rate of 30.09% and *P*<.001 (*t* test) when compared with 20 to 29 years-old group. The risk difference between the 50 to 59 years-old group and the 20 to 29 years-old group is 9.3% (SE 0.00059). Generally, the incidence rate increases with age in the nonpediatric groups (aged 30 years and younger). However, the groups of patients aged older than 60 years see a small drop in adverse event incidence, but the exact reason for this is still not clear. In general, Figure 1 shows a nonlinear trend appearance with peaks at the 0 to 9 years-old and 50 to 59 years-old groups.

Figure 2 lists the top adverse event examples in each of the age groups that show higher incidence rate across different clinical trials when compared to other age groups. Within each age group, we selected the top events that show significance (*P*<.01, *t* test) when compared with the comparison group. The comparison group consists of trials that reported the same event but with patients who were not in the same age group. For example, given the 0 to 9 years-old group and adverse event pharyngitis, we can find 316 trials that have an average incidence rate of 3.93% for the age group. The comparison group includes 299 trials that reported the same event among 10 to 100 years-old patient groups (at an average incidence rate of 2.17%) . Using *t* tests to compare both groups, we have *P*<.01, which statistically shows that pharyngitis is significantly higher among the young child group across clinical trials. The results in Figure 2 indicate that individual adverse events can have a significant disparity in term of incidence rate across different age groups. Commonly shared nonserious events are filtered (see Multimedia Appendix 1). The data also indicate there are strong patterns of adverse events in each of the age groups. For example, the 0 to 9 years-old group shows a significant number of adverse events in infection and infestation: 7 out of the top 9 events in the group are infection events. Young adults (20-29 years-old group) show adverse events with the reproductive system and musculoskeletal system; older adults (30-49 years-old group) show a higher level of adverse events in psychiatric and respiratory disorder categories. Blood system events and gastrointestinal events are higher in the 50 to 59 years-old group, and the oldest patients (60-100 years-old group) generally are at significantly higher risk for cardiac and vascular disorders than other age groups.

XSL•FO
RenderX

Figure 1. Adverse event incidence rate with different age groups. (X-axis: age group; Y-axis: micro-average of adverse event incidence in an age group; confidence intervals are shown on the bar.).

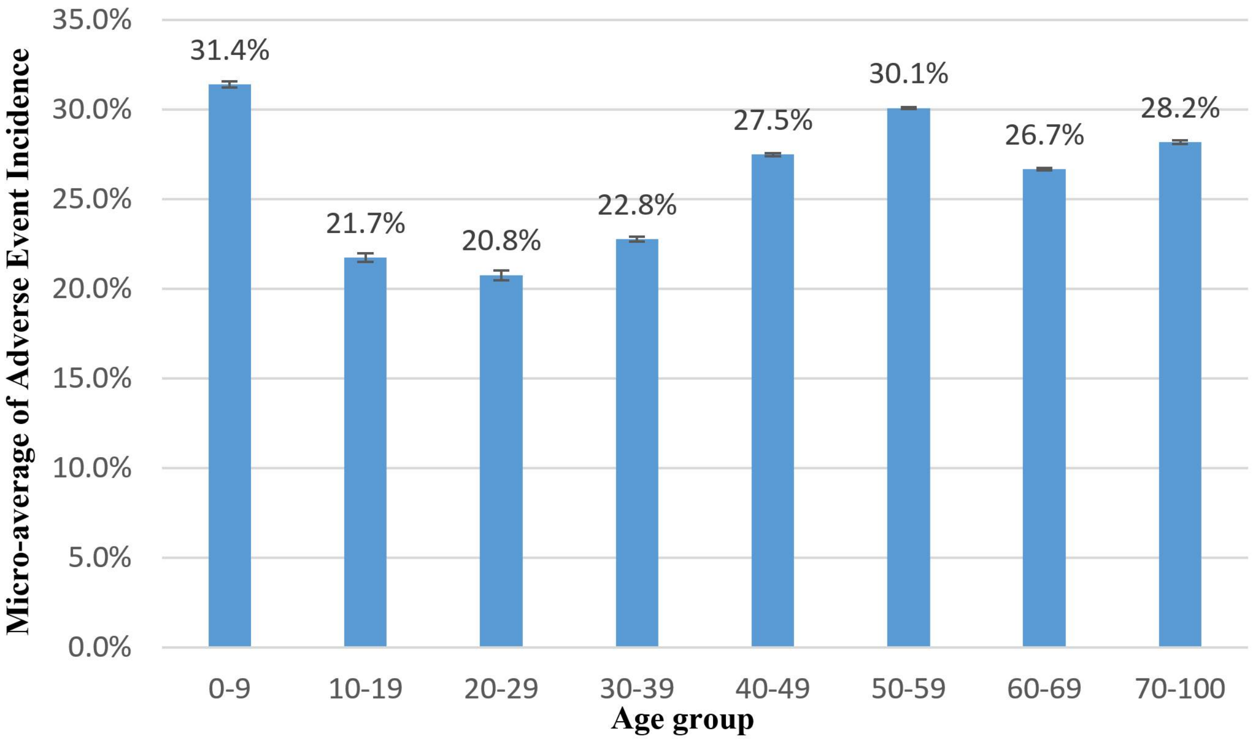
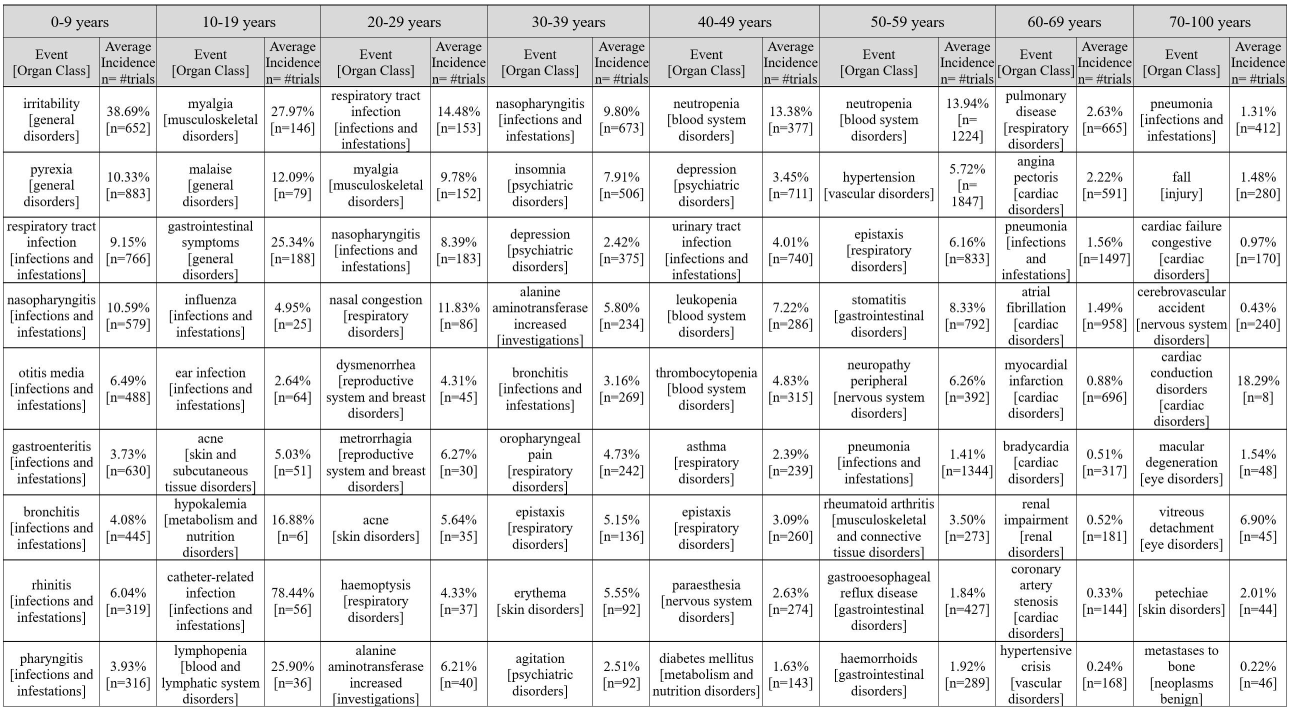


Figure 2. Top significant adverse event examples across clinical trials in each age group (*P*<.01). Shared nonserious events were filtered out.



0-9 years		10-19 years		20-29 years		30-39 years		40-49 years		50-59 years		60-69 years		70-100 years	
Event [Organ Class]	Average Incidence n= #trials	Event [Organ Class]	Average Incidence n= #trials	Event [Organ Class]	Average Incidence n= #trials	Event [Organ Class]	Average Incidence n= #trials	Event [Organ Class]	Average Incidence n= #trials	Event [Organ Class]	Average Incidence n= #trials	Event [Organ Class]	Average Incidence n= #trials	Event [Organ Class]	Average Incidence n= #trials
irritability [general disorders]	38.69% [n=652]	myalgia [musculoskeletal disorders]	27.97% [n=146]	respiratory tract infection [infections and infestations]	14.48% [n=153]	nasopharyngitis [infections and infestations]	9.80% [n=673]	neutropenia [blood system disorders]	13.38% [n=377]	neutropenia [blood system disorders]	13.94% [n= 1224]	pulmonary disease [respiratory disorders]	2.63% [n=665]	pneumonia [infections and infestations]	1.31% [n=412]
pyrexia [general disorders]	10.33% [n=883]	malaise [general disorders]	12.09% [n=79]	myalgia [musculoskeletal disorders]	9.78% [n=152]	insomnia [psychiatric disorders]	7.91% [n=506]	depression [psychiatric disorders]	3.45% [n=711]	hypertension [vascular disorders]	5.72% [n= 1847]	angina pectoris [cardiac disorders]	2.22% [n=591]	fall [injury]	1.48% [n=280]
respiratory tract infection [infections and infestations]	9.15% [n=766]	gastrointestinal symptoms [general disorders]	25.34% [n=188]	nasopharyngitis [infections and infestations]	8.39% [n=183]	depression [psychiatric disorders]	2.42% [n=375]	urinary tract infection [infections and infestations]	4.01% [n=740]	epistaxis [respiratory disorders]	6.16% [n=833]	pneumonia [infections and infestations]	1.56% [n=1497]	cardiac failure congestive [cardiac disorders]	0.97% [n=170]
nasopharyngitis [infections and infestations]	10.59% [n=579]	influenza [infections and infestations]	4.95% [n=25]	nasal congestion [respiratory disorders]	11.83% [n=86]	alanine aminotransferase increased [investigations]	5.80% [n=234]	leukopenia [blood system disorders]	7.22% [n=286]	stomatitis [gastrointestinal disorders]	8.33% [n=792]	atrial fibrillation [cardiac disorders]	1.49% [n=958]	cerebrovascular accident [nervous system disorders]	0.43% [n=240]
otitis media [infections and infestations]	6.49% [n=488]	ear infection [infections and infestations]	2.64% [n=64]	dysmenorrhea [reproductive system and breast disorders]	4.31% [n=45]	bronchitis [infections and infestations]	3.16% [n=269]	thrombocytopenia [blood system disorders]	4.83% [n=315]	neuropathy peripheral [nervous system disorders]	6.26% [n=392]	myocardial infarction [cardiac disorders]	0.88% [n=696]	cardiac conduction disorders [cardiac disorders]	18.29% [n=8]
gastroenteritis [infections and infestations]	3.73% [n=630]	acne [skin and subcutaneous tissue disorders]	5.03% [n=51]	metrorrhagia [reproductive system and breast disorders]	6.27% [n=30]	oropharyngeal pain [respiratory disorders]	4.73% [n=242]	asthma [respiratory disorders]	2.39% [n=239]	pneumonia [infections and infestations]	1.41% [n=1344]	bradycardia [cardiac disorders]	0.51% [n=317]	macular degeneration [eye disorders]	1.54% [n=48]
bronchitis [infections and infestations]	4.08% [n=445]	hypokalemia [metabolism and nutrition disorders]	16.88% [n=6]	acne [skin disorders]	5.64% [n=35]	epistaxis [respiratory disorders]	5.15% [n=136]	epistaxis [respiratory disorders]	3.09% [n=260]	rheumatoid arthritis [musculoskeletal and connective tissue disorders]	3.50% [n=273]	renal impairment [renal disorders]	0.52% [n=181]	vitreous detachment [eye disorders]	6.90% [n=45]
rhinitis [infections and infestations]	6.04% [n=319]	catheter-related infection [infections and infestations]	78.44% [n=56]	haemoptysis [respiratory disorders]	4.33% [n=37]	erythema [skin disorders]	5.55% [n=92]	paraesthesia [nervous system disorders]	2.63% [n=274]	gastrooesophageal reflux disease [gastrointestinal disorders]	1.84% [n=427]	coronary artery stenosis [cardiac disorders]	0.33% [n=144]	petechiae [skin disorders]	2.01% [n=44]
pharyngitis [infections and infestations]	3.93% [n=316]	lymphopenia [blood and lymphatic system disorders]	25.90% [n=36]	alanine aminotransferase increased [investigations]	6.21% [n=40]	agitation [psychiatric disorders]	2.51% [n=92]	diabetes mellitus [metabolism and nutrition disorders]	1.63% [n=143]	haemorrhoids [gastrointestinal disorders]	1.92% [n=289]	hypertensive crisis [vascular disorders]	0.24% [n=168]	metastases to bone [neoplasms benign]	0.22% [n=46]

Diversity of Adverse Events

Approximately 11,000 distinct adverse event types were observed in 6,808,619 participants. The adverse event diversity analysis was performed on the trial arm level, in which a group of patients received the same clinical intervention (eg, drug, surgery). We first analyzed the diversity among different groups of patients. Figure 3 shows that older groups of patients (aged 50 years and older) have a much higher diversity level of adverse events compared with the younger groups. The lowest diversity group is the 20 to 29 years-old young adult group. The group of young children (0-9 years-old group) also showed higher adverse event diversity than the young adult (20-29 years-old) group. On average, the young adult group observed 17.71 events/arm (95% CI 15.72-19.70, SE 1.02) of distinct adverse events. The young children group showed 32.58 events/arm (95% CI 31.49-34.71, SE 1.09) of distinct event types on average, which is approximately 1.84 times greater on average

XSL•FO
RenderX

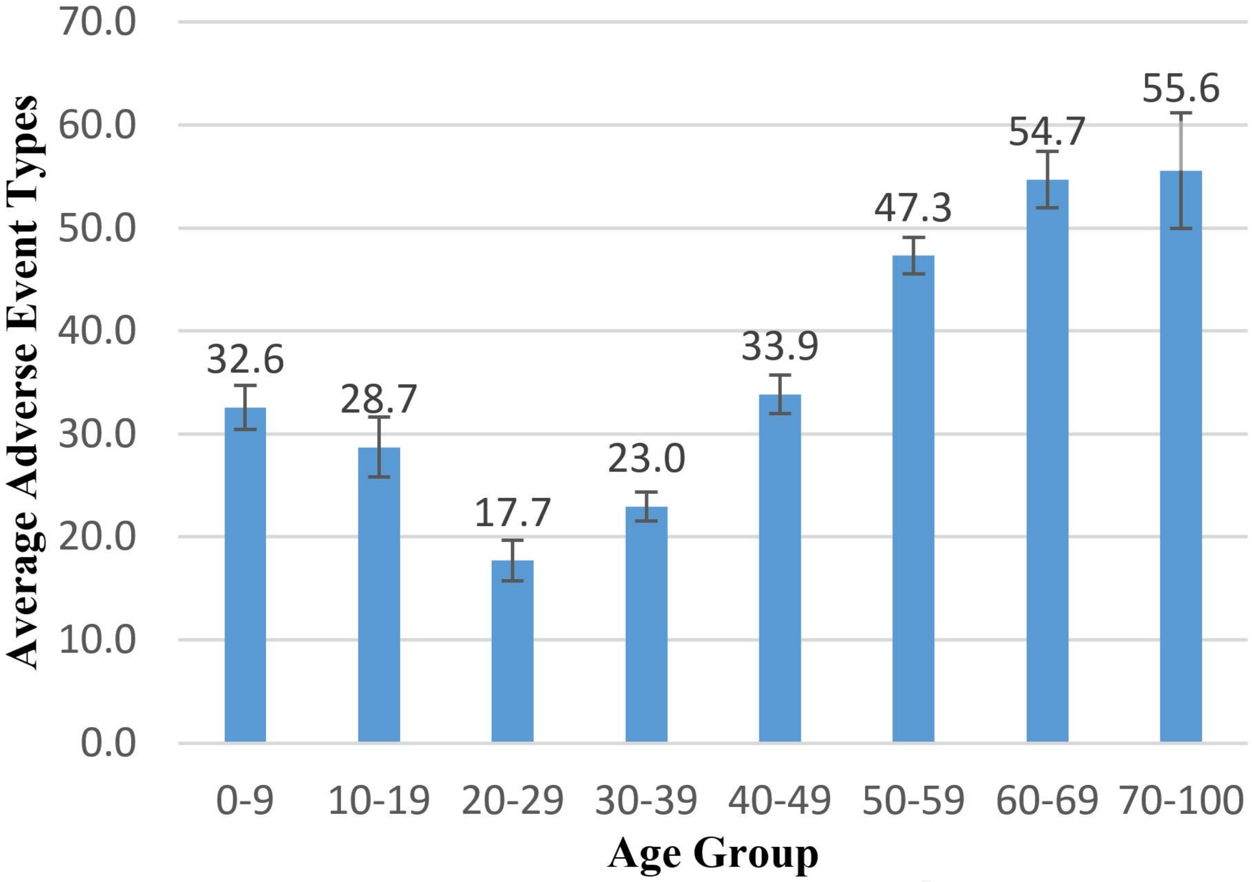
than the young adult group (*P*<.001). In comparison to the lowest affected 20 to 29 years-old group, the adverse event diversities of patients aged 30 to 69 years old increased significantly at an average rate of 77% for each age group as the patient age increased. The group aged 70 to 100 years showed the highest diversity level of 55.55 events/arm (95% CI 49.93-61.17, SE 2.867), which is approximately 3.44 times greater than the 20 to 29 years-old young adult group (*P*<.001). Clinical trials that recruited older patients showed significantly higher levels of adverse event diversity, and clinical trials with children younger than 20 years old also have a higher level of adverse event diversity in comparison to younger adults.

In Figure 4, the adverse events were classified into the 26 MedDRA SOCs. We analyzed the adverse event diversity in each of the age groups and SOCs. Note that event diversity values with low trial supports are shown in brackets; these events were documented in less than 30 clinical trials. Figure 4 displays a heat map of the results of the SOCs diversity analysis. Adverse event diversity is compared across the age groups in different organ categories in the same row. Higher diversity in the same category (ie, on same data row) is shown in red; lower diversity in green. The color intensity is rendered according to the percentile of diversity value when compared to the highest or lowest value. The overall pattern is similar to the previous analysis in which older patient groups (aged 50 years and older) generally showed more types of adverse events. However, when analyzing the diversity level in individual SOCs, we can observe some distinct patterns across the age groups. For example, the 0 to 9 years-old group has a high diversity level of adverse events in infections and infestations (10.36 events/arm), general disorders (6.16 events/arm), and skin and subcutaneous tissue disorders (5.21 events/arm) when compared to young adult group. The adverse events patients in the 10 to 19 years-old group are more likely to experience include ear and labyrinth disorder (1.94 events/arm); immune system disorders (1.75 events/arm); and pregnancy, puerperium, and perinatal conditions (3.79 events/arm) when compared to all other groups. The 20 to 29 years-old group is more diverse in congenital, familial, and genetic disorders (3.05 events/arm) and reproductive system and breast diseases (2.68 events/arm) and higher in pregnancy, puerperium, and perinatal conditions (4.98 events/arm). The 30 to 39 years-old group is more diverse in congenital, familial, and genetic disorders (2.32 events/arm) and pregnancy disorders (2.40 events/arm) and notably higher in psychiatric disorders (3.75 events/arm). The 40 to 49 years-old group also has a high level of event diversity in psychiatric disorders (3.56 events/arm). The three groups of patients older than 50 years generally have a higher event diversity than younger groups, except in the SOCs immune system; congenital, familial, and genetic disorders; and, as expected, pregnancy conditions.

Figure 5 shows the ranking of adverse event diversity in each of the age groups. The ranking is compared in the same age group (ie, same column) and ranked from 1 with highest diversity to 26 with lowest diversity across the 26 categories. The last column on the right shows the total rank of each category across all age groups. In each column, highest diversity value is shown in red and lowest is shown in white. Other values are rendered according their normalized value percentile between the highest and lowest value. The total results in Figure 5 show that the infection and infestations category has the highest average diversity level across most of the age groups, followed by gastrointestinal disorders and general disorders. The lowest categories are immune system disorders, endocrine disorders, and social circumstances.

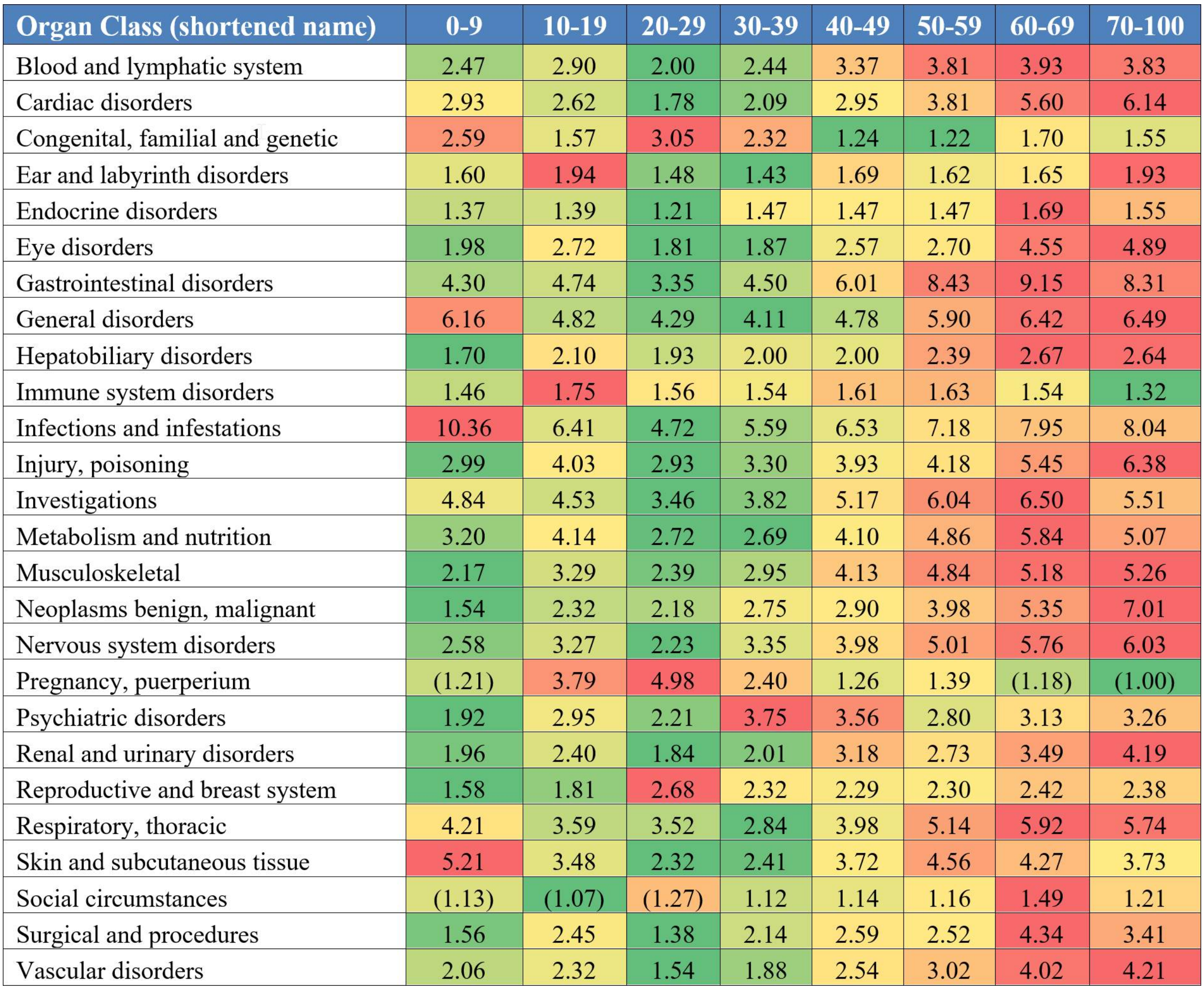
XSL•FO
RenderX

Figure 3. Diversity of adverse events among different age groups. (X-axis: age groups, Y-axis: average adverse event types; confidence intervals are shown on the bar.).



XSL•FO
RenderX

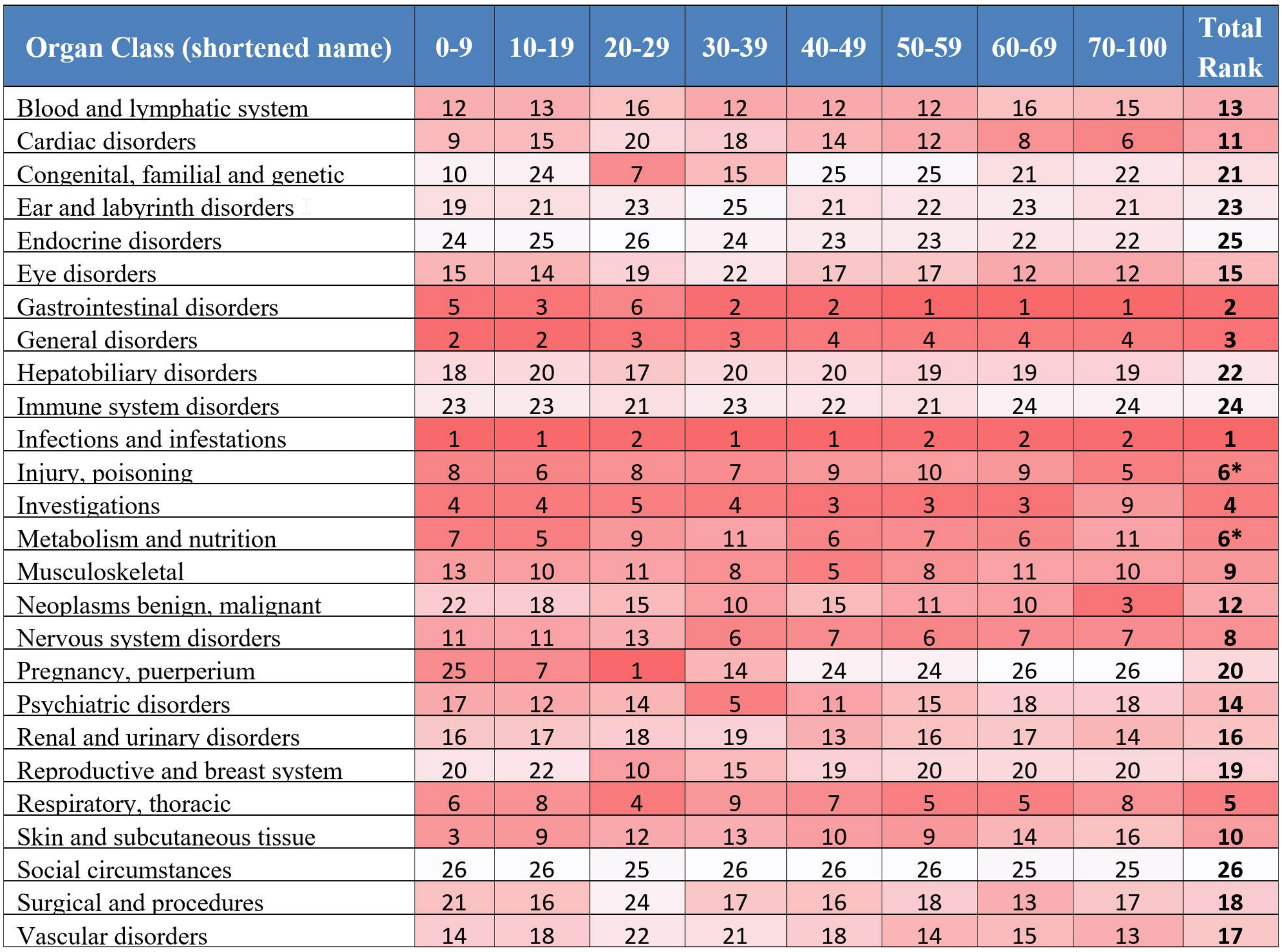
Figure 4. Average adverse event diversity in Medical Dictionary for Regulatory Activities (MedDRA) organ classes across different age groups.



Organ Class (shortened name)	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-100
Blood and lymphatic system	2.47	2.90	2.00	2.44	3.37	3.81	3.93	3.83
Cardiac disorders	2.93	2.62	1.78	2.09	2.95	3.81	5.60	6.14
Congenital, familial and genetic	2.59	1.57	3.05	2.32	1.24	1.22	1.70	1.55
Ear and labyrinth disorders	1.60	1.94	1.48	1.43	1.69	1.62	1.65	1.93
Endocrine disorders	1.37	1.39	1.21	1.47	1.47	1.47	1.69	1.55
Eye disorders	1.98	2.72	1.81	1.87	2.57	2.70	4.55	4.89
Gastrointestinal disorders	4.30	4.74	3.35	4.50	6.01	8.43	9.15	8.31
General disorders	6.16	4.82	4.29	4.11	4.78	5.90	6.42	6.49
Hepatobiliary disorders	1.70	2.10	1.93	2.00	2.00	2.39	2.67	2.64
Immune system disorders	1.46	1.75	1.56	1.54	1.61	1.63	1.54	1.32
Infections and infestations	10.36	6.41	4.72	5.59	6.53	7.18	7.95	8.04
Injury, poisoning	2.99	4.03	2.93	3.30	3.93	4.18	5.45	6.38
Investigations	4.84	4.53	3.46	3.82	5.17	6.04	6.50	5.51
Metabolism and nutrition	3.20	4.14	2.72	2.69	4.10	4.86	5.84	5.07
Musculoskeletal	2.17	3.29	2.39	2.95	4.13	4.84	5.18	5.26
Neoplasms benign, malignant	1.54	2.32	2.18	2.75	2.90	3.98	5.35	7.01
Nervous system disorders	2.58	3.27	2.23	3.35	3.98	5.01	5.76	6.03
Pregnancy, puerperium	(1.21)	3.79	4.98	2.40	1.26	1.39	(1.18)	(1.00)
Psychiatric disorders	1.92	2.95	2.21	3.75	3.56	2.80	3.13	3.26
Renal and urinary disorders	1.96	2.40	1.84	2.01	3.18	2.73	3.49	4.19
Reproductive and breast system	1.58	1.81	2.68	2.32	2.29	2.30	2.42	2.38
Respiratory, thoracic	4.21	3.59	3.52	2.84	3.98	5.14	5.92	5.74
Skin and subcutaneous tissue	5.21	3.48	2.32	2.41	3.72	4.56	4.27	3.73
Social circumstances	(1.13)	(1.07)	(1.27)	1.12	1.14	1.16	1.49	1.21
Surgical and procedures	1.56	2.45	1.38	2.14	2.59	2.52	4.34	3.41
Vascular disorders	2.06	2.32	1.54	1.88	2.54	3.02	4.02	4.21

XSL•FO
RenderX

Figure 5. Ranking of average adverse event diversity in each of the age groups. The rankings are calculated within each group. Two organ classes are tied at the sixth place in the total rank (marked with asterisk).



Organ Class (shortened name)	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-100	Total Rank
Blood and lymphatic system	12	13	16	12	12	12	16	15	**13**
Cardiac disorders	9	15	20	18	14	12	8	6	**11**
Congenital, familial and genetic	10	24	7	15	25	25	21	22	**21**
Ear and labyrinth disorders	19	21	23	25	21	22	23	21	**23**
Endocrine disorders	24	25	26	24	23	23	22	22	**25**
Eye disorders	15	14	19	22	17	17	12	12	**15**
Gastrointestinal disorders	5	3	6	2	2	1	1	1	**2**
General disorders	2	2	3	3	4	4	4	4	**3**
Hepatobiliary disorders	18	20	17	20	20	19	19	19	**22**
Immune system disorders	23	23	21	23	22	21	24	24	**24**
Infections and infestations	1	1	2	1	1	2	2	2	**1**
Injury, poisoning	8	6	8	7	9	10	9	5	**6***
Investigations	4	4	5	4	3	3	3	9	**4**
Metabolism and nutrition	7	5	9	11	6	7	6	11	**6***
Musculoskeletal	13	10	11	8	5	8	11	10	**9**
Neoplasms benign, malignant	22	18	15	10	15	11	10	3	**12**
Nervous system disorders	11	11	13	6	7	6	7	7	**8**
Pregnancy, puerperium	25	7	1	14	24	24	26	26	**20**
Psychiatric disorders	17	12	14	5	11	15	18	18	**14**
Renal and urinary disorders	16	17	18	19	13	16	17	14	**16**
Reproductive and breast system	20	22	10	15	19	20	20	20	**19**
Respiratory, thoracic	6	8	4	9	7	5	5	8	**5**
Skin and subcutaneous tissue	3	9	12	13	10	9	14	16	**10**
Social circumstances	26	26	25	26	26	26	25	25	**26**
Surgical and procedures	21	16	24	17	16	18	13	17	**18**
Vascular disorders	14	18	22	21	18	14	15	13	**17**

Discussion

Principal Findings

We conducted a population study to analyze the adverse event risk among clinical trial participants. This study differs from patient-level adverse event analysis in that we integrated large amounts of clinical trial data to conduct a population-level analysis looking at adverse event risk patterns across different age groups. We found that young pediatric patients and older patients have a higher level of incidence and diversity of adverse events. The total incidence of adverse events in the youngest age group is higher compared with all other groups. Additionally, the incidence rate of adverse events in this group is significantly higher in the infectious event and general event categories. The older adult groups (aged older than 60 years) showed a comparatively higher incidence of cardiac disorders and vascular disorders. When compared across the 26 SOCs, we observed that the diversity of adverse event patterns differs significantly across the age groups. Older patients show a significantly higher level of adverse event diversity in most of the SOCs, while the younger age groups show higher levels within some SOCs.

Related Studies

Previous studies have focused on the incidence of adverse events in population levels in various clinical settings. The Canadian Adverse Events Study [18] reported an adverse event rate of 7.5% in 2.8 million hospital admissions. Older patients were more likely to be affected by adverse events. The study also suggests that 9250 to 23,750 deaths from adverse events could have been prevented among the 2.5 million admissions to acute-care hospitals in Canada. A study on 1000 discharged patient records showed that elderly patients (aged 65 years and older) had a high incidence of adverse drug events (18.7%) [19]. Among the identified events, 35% were considered preventable and 32% were serious events. A systematic review of 8 studies [20] on in-hospital adverse events in 6 countries shows that the median incidence of adverse events was 9.2%, and about 43.5% of the adverse events could be preventable. In the outpatient setting, a study showed that adverse event–related visits increased between 1995 and 2005 [21]. Furthermore, the incidence of adverse events also increases with patient age. This study indicated that patient age was one of the important risk factors for adverse event–related visits. Patients aged 65 years and older had a peak of adverse event visits of 47 per 1000 patients. A pediatric study [22] showed that adverse events occurred in about 1% of the pediatric hospitalizations, of which about 0.6% were preventable events compared with a rate of 1.5% in nonelderly adults. The Critical Care Safety Study [23] showed that among 391 studied patients, 20.2% were affected by 120 adverse events and 54% of the events were preventable.

XSL•FO
RenderX

Compared to these studies which focused on preventable adverse events in health care settings, the adverse event rate in clinical studies is significantly higher in terms of incidence rates in all age groups at an average of 27.0%. Many clinical study interventions are experimental in nature and thus are associated inherently with a higher level of risk than normal clinical interventions. In-hospital treatments normally use matured intervention protocols that use validated postmarketing drugs or procedures, whereas clinical trials are often designed to test experimental interventions. For example, in clinical trials aimed to develop new drugs, only about 1 in 10 will be approved by the US Food and Drug Administration [24,25]. Many trials are canceled in the process or the tested substance is disapproved due to risk of adverse events. This suggests that adverse event risk estimation is critical for clinical study preparation. This study provides a quantitative reference for clinical investigators to estimate the trial adverse event risk for targeted age groups when planning clinical trials.

Clinical Trial Adverse Events and Participant Age

Age is one of the most commonly used clinical study recruitment criteria [26,27], and the risk for adverse events is a primary criterion for evaluating the safety of the targeted intervention in a clinical study [28]. However, few systematic studies have explored the association between adverse clinical trial outcomes and participant age. This study fills the gap by focusing on the adverse event patterns in clinical trials at the population-health level. This study shows that age-related adverse events could be an important factor for clinical trial planning, recruitment, and monitoring. Furthermore, the importance of recruiting more children in clinical trials has been discussed in various reviews [29,30]. The risk of adverse events in children is higher, as suggested by our study results; however, even though numerous regulations have been established to improve children's safety in clinical studies, there is still a lack of evidence-based support to help clinical investigators estimate the adverse event risks for children at the early stages of a clinical study [28,31]. This study suggests that the adverse event distribution shows strong categorical patterns among age groups, providing a population baseline for estimating the risk of adverse events. Similarly, many studies have verified that older patients have a higher risk of adverse events. Our study shows that among older populations, not only is the adverse event incidence rate higher, the diversity of adverse events also is significantly higher in clinical trials. Furthermore, specific adverse events may be more common in one age group compared to another as seen with the higher incidence of infectious events in the young children group or the peak of psychiatric disorders in the middle age group.

Limitations and Future Work

This study is limited due to the data granularity on ClinicalTrials.gov. The report on ClinicalTrials.gov does not include adverse event diversity at the individual patient level; for example, we cannot determine how many different adverse events occurred in an individual patient. Therefore, we performed the adverse event diversity analysis on the trial arm level and categorized events by the MedDRE organ classes. The inability to identify individual patients may also create bias when a patient joins multiple trials, although we estimate the proportion of patients joining multiple trials is low because most trials exclude patients who are participating in other trials concurrently. Furthermore, certain types of studies may be more common in one age group than another which could lead to a higher incidence of a type of adverse event. For instance, perhaps few psychiatric studies are performed in the younger patients in comparison to the older patients. For nonserious events, some trials on ClinicalTrials.gov only reported events that exceeded a frequency of 5% within any arm of the trials. This could lead to potential undercount of nonserious events. We used MetaMap [14] to normalized terminologies, which may not normalize terms 100% correctly to the UMLS concepts. However, a few studies evaluated the performance of MetaMap [32,33] and found that the accuracy of MetaMap was over 90%. The MedDRA system classes were updated in March 2016 to include a new category called product issues. The new system class contains events related to device issues. We currently have no adverse events mapped to this category. We also want to compare the differences of adverse event patterns between the intervention groups and the placebo groups on the population level. However, it requires us to develop new natural language processing methods to systematically identify placebo and intervention arms from the free-text trial arm descriptions. This will be our future work.

Conclusions

The adverse event incidence rate in clinical trial studies is as high as 27.0% at the population level, which is higher than the reported incident rate in various patient care settings (7%-20%). Clinical trials may include a greater risk in terms of adverse events by their nature. Young children and older patients have higher risks of adverse events in clinical trials. The pattern of adverse event types in different organ categories is different across the age groups. Evidence-based risk analysis should be used to facilitate clinical trial design and planning.

Acknowledgments

The study received funding from the University of Wisconsin-Milwaukee College of Health Sciences' Stimulus for Extramural Enhancement and Development Fund. The corresponding author (JL) had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

XSL•FO
RenderX

Authors' Contributions

JL conceived and designed the study. RAC and JL contributed to the structure of the paper. JL collected the data. JL, CE, CC, and RAC provided interpretation of results and contributed to the writing of the manuscript. RAC, CC, and CE revised the content critically. JL, CE, CC, and RAC reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Common adverse events that were filtered out.

[XLSX File (Microsoft Excel File), 8KB - medinform_v4i4e30_app1.xlsx]

References

1. ClinicalTrials.gov. 2016. URL: https://clinicaltrials.gov/ [accessed 2016-10-02] [WebCite Cache ID 6kyELH9ol]
2. Lazarou J, Pomeranz B, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. JAMA 1998 Apr 15;279(15):1200-1205. [Medline: 9555760]
3. Kaushal R, Bates D, Landrigan C, McKenna K, Clapp M, Federico F, et al. Medication errors and adverse drug events in pediatric inpatients. JAMA 2001 Apr 25;285(16):2114-2120. [Medline: 11311101]
4. Moore T, Weiss S, Kaplan S, Blaisdell CJ. Reported adverse drug events in infants and children under 2 years of age. Pediatrics 2002 Nov;110(5):e53. [Medline: 12415059]
5. Rich M, McSherry F, Williford W, Yusuf S, Digitalis Investigation Group. Effect of age on mortality, hospitalizations and response to digoxin in patients with heart failure: the DIG study. J Am Coll Cardiol 2001 Sep;38(3):806-813 [FREE Full text] [Medline: 11527638]
6. Huddleston J, Wang Y, Uquillas C, Herndon J, Maloney WJ. Age and obesity are risk factors for adverse events after total hip arthroplasty. Clin Orthop Relat Res 2012 Feb;470(2):490-496 [FREE Full text] [doi: 10.1007/s11999-011-1967-y] [Medline: 21796477]
7. Roland NJ, Bhalla RK, Earis J. The local side effects of inhaled corticosteroids: current understanding and review of the literature. Chest 2004 Jul;126(1):213-219. [doi: 10.1378/chest.126.1.213] [Medline: 15249465]
8. Zarin D, Tse T, Williams R, Califf R, Ide NC. The ClinicalTrials.gov results database: update and key issues. N Engl J Med 2011 Mar 3;364(9):852-860 [FREE Full text] [doi: 10.1056/NEJMsa1012065] [Medline: 21366476]
9. Luo Z, Zhang GQ, Xu R. Mining patterns of adverse events using aggregated clinical trial results. AMIA Jt Summits Transl Sci Proc 2013;2013:112-116 [FREE Full text] [Medline: 24303317]
10. Brown E, Wood L, Woods S. The medical dictionary for regulatory activities (MedDRA). Drug Safety 1999:109-117.
11. Trotti A, Colevas A, Setser A, Basch E. Patient-reported outcomes and the evolution of adverse event reporting in oncology. J Clin Oncol 2007:5121-5127.
12. Edwards I, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. Lancet 2000;356:1255-1259.
13. International Classification of Diseases, Ninth Revision, Clinical Modification: Geneva, Switzerland: World Health Organization; 1996.
14. 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]
15. Aronson A, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
16. Vincent C, Neale G, Woloshynowych M. Adverse events in British hospitals: preliminary retrospective record review. BMJ 2001:322-517.
17. Classen D, Pestotnik S, Evans R, Lloyd J, Burke JP. Adverse drug events in hospitalized patients and excess length of stay, extra costs, and attributable mortality. JAMA 1997:301-306. [Medline: 9002492]
18. Baker GR, 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]
19. Kanaan A, Donovan J, Duchin N, Field T, Tjia J, Cutrona SL, et al. Adverse drug events after hospital discharge in older adults: types, severity, and involvement of Beers Criteria Medications. J Am Geriatr Soc 2013 Nov;61(11):1894-1899 [FREE Full text] [doi: 10.1111/jgs.12504] [Medline: 24116689]
20. de Vries EN, Ramrattan M, Smorenburg S, Gouma D, Boermeester MA. The incidence and nature of in-hospital adverse events: a systematic review. Qual Saf Health Care 2008 Jun;17(3):216-223 [FREE Full text] [doi: 10.1136/qshc.2007.023622] [Medline: 18519629]

XSL•FO
RenderX

21. Bourgeois FT, Shannon MW, Valim C, Mandl KD. Adverse drug events in the outpatient setting: an 11-year national analysis. Pharmacoepidemiol Drug Saf 2010 Sep;19(9):901-910 [FREE Full text] [doi: 10.1002/pds.1984] [Medline: 20623513]
22. Woods D, Thomas E, Holl J, Altman S, Brennan T. Adverse events and preventable adverse events in children. Pediatrics 2005 Jan;115(1):155-160. [doi: 10.1542/peds.2004-0410] [Medline: 15629994]
23. Rothschild J, Landrigan C, Cronin J, Kaushal R, Lockley S, Burdick E, et al. The Critical Care Safety Study: The incidence and nature of adverse events and serious medical errors in intensive care. Crit Care Med 2005 Aug;33(8):1694-1700. [Medline: 16096443]
24. Hay M, Thomas D, Craighead J, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol 2014 Jan;32(1):40-51. [doi: 10.1038/nbt.2786] [Medline: 24406927]
25. DiMasi J, Feldman L, Seckler A, Wilson A. Trends in risks associated with new drug development: success rates for investigational drugs. Clin Pharmacol Ther 2010 Mar;87(3):272-277. [doi: 10.1038/clpt.2009.295] [Medline: 20130567]
26. Luo Z, Yetisgen-Yildiz M, Weng C. Dynamic categorization of clinical research eligibility criteria by hierarchical clustering. J Biomed Inform 2011 Dec;44(6):927-935 [FREE Full text] [doi: 10.1016/j.jbi.2011.06.001] [Medline: 21689783]
27. Luo Z, Johnson S, Weng C. Semi-automatically inducing semantic classes of clinical research eligibility criteria using UMLS and hierarchical clustering. AMIA Annu Symp Proc 2010;2010:487-491 [FREE Full text] [Medline: 21347026]
28. Singh S, Loke YK. Drug safety assessment in clinical trials: methodological challenges and opportunities. Trials 2012;13:138 [FREE Full text] [doi: 10.1186/1745-6215-13-138] [Medline: 22906139]
29. Caldwell PH, Murphy S, Butow P, Craig JC. Clinical trials in children. Lancet 2004;364(9436):803-811. [doi: 10.1016/S0140-6736(04)16942-0] [Medline: 15337409]
30. Knox C, Burkhart PV. Issues related to children participating in clinical research. J Pediatr Nurs 2007 Aug;22(4):310-318. [doi: 10.1016/j.pedn.2007.02.004] [Medline: 17645958]
31. Shaddy R, Denne SC, Committee on Pediatric Research. Clinical report: guidelines for the ethical conduct of studies to evaluate drugs in pediatric populations. Pediatrics 2010 Apr;125(4):850-860. [doi: 10.1542/peds.2010-0082] [Medline: 20351010]
32. Pratt W, Yetisgen-Yildiz M. A study of biomedical concept identification: MetaMap versus people. AMIA Annu Symp Proc 2003:529-533 [FREE Full text] [Medline: 14728229]
33. Osborne J, Gyawali B, Solorio T. Evaluation of YTEX and MetaMap for clinical concept recognition. arXiv 2014;arXiv:1402.1668.

Abbreviations

MedDRA: Medical Dictionary for Regulatory Activities
SOC: system organ classes
UMLS: Unified Medical Language System

Edited by CL Parra-Calderón; submitted 02.08.16; peer-reviewed by S Van Poucke, L Cui; comments to author 22.08.16; revised version received 23.09.16; accepted 30.09.16; published 17.10.16.

Please cite as:
Luo J, Eldredge C, Cho CC, Cisler RA
Population Analysis of Adverse Events in Different Age Groups Using Big Clinical Trials Data
JMIR Med Inform 2016;4(4):e30
URL: http://medinform.jmir.org/2016/4/e30/
doi:10.2196/medinform.6437
PMID:27751983

©Jake Luo, Christina Eldredge, Chi C Cho, Ron A Cisler. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 17.10.2016. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Challenges and Opportunities of Big Data in Health Care: A Systematic Review

Clemens Scott Kruse[1], MBA, MHA, MSIT, PhD; Rishi Goswamy[1], MHA; Yesha Raval[1], MHA; Sarah Marawi[1], MHA

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

Corresponding Author:
Clemens Scott Kruse, MBA, MHA, MSIT, PhD
School of Health Administration
Texas State University
601 University Dr
College of Health Professions
San Marcos, TX, 78666
United States
Phone: 1 2103554742
Fax: 1 5122458712
Email: scottkruse@txstate.edu

Abstract

Background: Big data analytics offers promise in many business sectors, and health care is looking at big data to provide answers to many age-related issues, particularly dementia and chronic disease management.

Objective: The purpose of this review was to summarize the challenges faced by big data analytics and the opportunities that big data opens in health care.

Methods: A total of 3 searches were performed for publications between January 1, 2010 and January 1, 2016 (PubMed/MEDLINE, CINAHL, and Google Scholar), and an assessment was made on content germane to big data in health care. From the results of the searches in research databases and Google Scholar (N=28), the authors summarized content and identified 9 and 14 themes under the categories *Challenges* and *Opportunities*, respectively. We rank-ordered and analyzed the themes based on the frequency of occurrence.

Results: The top challenges were issues of data structure, security, data standardization, storage and transfers, and managerial skills such as data governance. The top opportunities revealed were quality improvement, population management and health, early detection of disease, data quality, structure, and accessibility, improved decision making, and cost reduction.

Conclusions: Big data analytics has the potential for positive impact and global implications; however, it must overcome some legitimate obstacles.

(JMIR Med Inform 2016;4(4):e38) doi:10.2196/medinform.5359

KEYWORDS

big data; analytics; health care; human genome; electronic medical record

Introduction

Rationale

Big data analytics offers promise in many business sectors, and health care is looking at big data to provide answers to many age-related issues, particularly dementia and chronic disease management. This systematic review explores the depth of big data analytics since 2010 and identifies both challenges and opportunities associated with big data in health care. The review follows the standard set by Preferred Reporting Items for Systematic Reviews and Meta-analysis (2009) [1].

Big data is commonly defined through the 4 Vs: volume (scale or quantity of data), velocity (speed and analysis of real-time or near-real-time data), variety (different forms of data, often from disparate data sources), and veracity (quality assurance of the data). The first 3 Vs are found in most literature [2,3], and the fourth V is a goal [4].

As of 2012, about 2.5 exabytes of data are created each day; Walmart can collect up to 2.5 petabytes of customer-related

XSL•FO
RenderX

data per hour [2]. The industry of health care produces and collects data at a staggering speed, but different electronic health records (EHRs) collect data in different structures: structured, unstructured, and semistructured. This variety can pose difficulty when seeking veracity or quality assurance of the data. The EHRs can provide a rich source of data, ripe for analysis to increase our understanding of disease mechanisms, as well as better and personalized health care, but the data structures pose a problem to standard means of analysis [5].

There are several large sources for big data in health care: genomics, EHR, medical monitoring devices, wearable video devices, and health-related mobile phone apps. Approximately 483 studies on genomics are registered with the US Department of Health and Human Services; these studies are being conducted in 9 countries, and they all use portions of the data from the Human Genome Project [6]. The EHR, being adopted in many countries, offers a source of data the depth of which is almost inconceivable. About 500 petabytes of data was generated by the EHR in 2012, and by 2020, the data will reach 25,000 petabytes [7]. The EHR can collect data from other monitoring devices, but the continuous data streams are not consistently saved in the longitudinal record.

The decrease in the cost of storage has enabled an exponential distribution of data collection, but the ability to analyze this quantity of data is the center of gravity for “big data” in health care. In the United States, financial incentives offered for the “meaningful use” of health information technology has spurred growth in the adoption of the EHR and other enabling health-related technology since 2009.

Health information systems show great potential in improving the efficiency in the delivery of care, a reduction in overall costs to the health care system, as well as a marked increase in patient outcomes [8]. The US government has allocated billions of dollars to help the country’s health care market realize some of these efficiencies and savings. Specific provisions of the Health Information Technology for Economic and Clinical Health (HITECH), part of the American Recovery and Reinvestment Act, acknowledge the importance of IT in the delivery of health care within the United States [9]. The Act allocates approximately US $17.2 billion in incentives for the adoption and meaningful use of health information technology, part of which involves the participation in the electronic exchange of clinical information. In 2010, the Congress passed the Health Information Exchange (HIE) Challenge Grant Program, which contributed about US $547.7 million to state HIE programs [10].

With the implementation of this legislation as well as the technologies associated with it, it is imperative to effectively organize and process the ever-increasing quantity of data that is digitally collected and stored within health care organizations. Other industries such as astronomy, retail, search engines, and politics have developed advanced data-handling capabilities to convert data into knowledge. Health care needs to follow their lead so that decisions regarding organizational objectives and goals can be met [4,11,12]. This evolutionary process of data management is collectively known as big data, and it is essential to the future of adoption and management of health information technology [13].

Objectives

The purpose of this systematic review is to objectively review articles and studies published in academic journals in order to compile a list of challenges and opportunities faced by big data analytics in health care in the United States. Particular emphasis was paid to age-related applications of big data.

Methods

Eligibility Criteria

Articles and studies were eligible for analysis if they were published between 2010 and 2015, published in academic journals, and published in English. The researchers chose a range from 2010 to 2015 for two reasons: HITECH was passed in 2009, and it appeared that a blossom of research and other articles seemed to occur in 2010. We focused on academic journals for their peer-review quality and to decrease the chance of selecting something about big data published from a noncredible source.

Information Sources

A combination of key terms from Medical Subject Headings (MeSH) and Boolean operators were combined and used in 2 common research databases, CINAHL and PubMed, and combined with a general search from Google Scholar (see Figure 1) in January 2016.

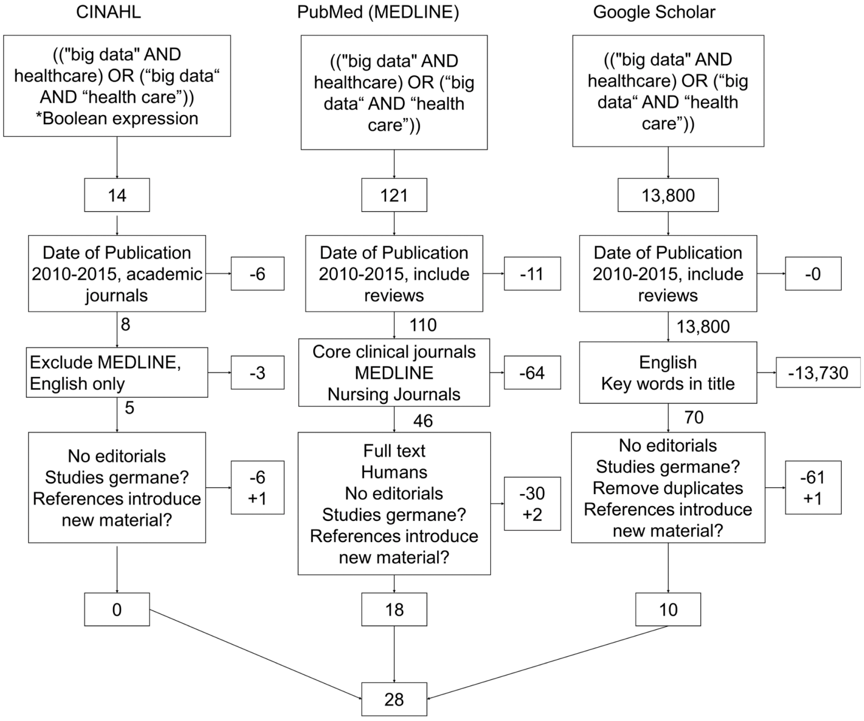
These terms were chosen not only because they are the focus of the review, but also because they were identified in the initial research into the definition of big data.

Search

The following search string was used in all 3 searches: ((“big data” AND healthcare) OR (“big data” AND “health care”)). This search string was used in CINAHL, PubMed (MEDLINE), and Google Scholar. In the 2 research databases, our team was able to restrict the search to academic journals (including other systematic reviews). MEDLINE was excluded in CINAHL because it was already captured in PubMed. Google Scholar creates difficulty for searches because of its severe limit of filters typically associated with academic research. The initial 13,935 results were limited by restricting dates to the last 5 years, limiting results to academic journals and MEDLINE, and in Google Scholar by restricting the keyword search to titles. The result from the filters ended with 121 articles to review.

XSL•FO
RenderX

Figure 1. Literature review process with inclusion and exclusion criteria.


CINAHL
PubMed (MEDLINE)
Google Scholar
(("big data" AND healthcare) OR ("big data" AND "health care")) *Boolean expression
(("big data" AND healthcare) OR ("big data" AND "health care"))
(("big data" AND healthcare) OR ("big data" AND "health care"))
14
121
13,800
Date of Publication 2010-2015, academic journals
-6
Date of Publication 2010-2015, include reviews
-11
Date of Publication 2010-2015, include reviews
-0
8
110
13,800
Exclude MEDLINE, English only
-3
Core clinical journals MEDLINE Nursing Journals
-64
English Key words in title
-13,730
5
46
70
No editorials Studies germane? References introduce new material?
-6 +1
Full text Humans No editorials Studies germane? References introduce new material?
-30 +2
No editorials Studies germane? Remove duplicates References introduce new material?
-61 +1
0
18
10
28

Study Selection

Through group research and a series of consensus meetings, researchers were trained to identify articles germane to this review and to recommend elimination of all others. A shared spreadsheet was used by the research team to parse through the list of articles. Researchers read all articles in their entirety. A total of 97 articles were eliminated due to various exclusion criteria (not germane to big data or health care, editorial only, not an academic journal, or duplicate from another search), and 4 additional articles were identified from the references of the 24 that remained. The group of reviewers made these rejections or additional recommendations through a series of consensus meetings where we met to discuss their recommendations and consensus was reached through discussion. A total of 28 articles remained in the final review.

Data Collection Process and Identification of Summary Measures

Each article was reviewed by at least two authors to identify the relevant points. All reviewers used a spreadsheet template to summarize their key observations from each article. One team member combined the spreadsheets into one and shared it once again. Reviewers held one more consensus meeting to discuss their findings. From this meeting, trends were identified, and from those trends, inferences were made.

Additional Analysis

From the list of observations, reviewers were able to identify some common threads that emerged as challenges and opportunities in health care that permeated multiple articles. Separate tables were created to group the threads, and from each of these tables, common themes were identified. These common themes only emerged when reviewers combined their observations. These themes were tabulated and counted for additional analysis.

Results

Study Selection

As depicted in Figure 1, 935 articles resulted from the initial search. Filters such as data published (2010-2015), academic journals, and English language were implemented to reduce the range to what was being studied. Reviewers agreed to eliminate editorials and focus on those articles that studied big data, as described in the Introduction section of this manuscript. At the end of the search process, only 28 remained. The articles reviewed for this study ranged from 2012 to 2015. The majority of the literature chosen for this paper was published in 2014 (15/28, 54%), and a minority was published in 2015 (2/28, 7%); the latter was most likely due to the early part of the year when the search was conducted.

Synthesis of Results

Multiple reviewers read each article in its entirety. Articles were included or excluded based on the criteria illustrated in Figure 1. All articles included in the analysis were sorted by date and are listed in Multimedia Appendix 1.

A study catalog number was assigned to each article to simplify the analysis. Researchers summarized the main points of each article for further analysis.

Additional Analysis

Through the combination of observations, reviewers identified common threads (challenges and opportunities) and themes from each thread. Themes were organized into affinity diagrams (Tables 1 and 2), compared, and discussed among researchers.

XSL•FO
RenderX

Challenges for Big Data in Health Care

Nine themes emerged under the category of challenges: data structure, security, data standardization, data storage and transfers, managerial issues such as governance and ownership, lack of skill of data analysts, inaccuracies in data, regulatory compliance, and real-time analytics. Examples for each theme are provided in Table 1. A total of 60 observations were made for challenges.

Table 1. Themes associated with challenges for big data in health care.

Themes	Examples	Number of articles (n)	Articles themes appeared in	% of total articles (N=28)
Data structure	Fragmented data	17	1, 2, 7-9, 12, 14-19, 22, 25-28	61%
	Incompatible formats			
	Heterogeneous data			
	Raw and unstructured datasets			
	Large volumes			
	High variety and velocity			
	Lack of transparency			
Security	Privacy	14	2, 4, 7-9, 12, 13, 17, 21, 22, 25-28	50%
	Confidentiality			
	Data duplication			
	Integrity			
Data standardization	Limited Interoperability	11	4, 5, 7-9, 11, 12, 15, 16, 22, 25	39%
	Data acquisition and cleansing			
	Global sharing			
	Terminology			
	Language barriers			
Storage and transfers	Expensive to store	8	1, 4, 7, 12, 22, 26, 28	28%
	Transfer from one place to other			
	Store electronic data			
	Securely extract, transmit, and process			
Managerial issues	Governance issues	4	2, 8, 14, 22	14%
	Ownership issues			
Lack of skill	Untrained workers	3	5, 9, 14	11%
Inaccuracies	Inconsistences	1	9	4%
	Lack of precision			
	Data timeliness			
Regulatory compliance	Legal concerns	1	13	4%
Real-time analytics	Real-time analytics	1	9	4%

The 4 Vs appear in multiple places under the Challenges category. Volume and variety are seen by name under the theme of Data structure. Variety is also implied in the same theme, but listed as Incompatible formats, as well as Raw and unstructured datasets. Variety can also be inferred from the theme of Data standardization, listed as Limited interoperability. Velocity is seen in the theme Real-time analytics. Veracity is seen under the theme of Data Standardization, but listed as Data acquisition and cleansing, Terminology, and Language barriers. It is also inferred in the theme Inaccuracies listed as Inconsistences and Lack of precision.

Data Structure Issues

Issues related to data structure were addressed in the majority of the papers reviewed for this study. It is essential that the key functions of data processing are supported by the applications of big data [13]. Big data applications should be user-friendly, transparent, and menu-driven [13,14]. The majority of data in health care is unstructured, such as from natural language processing [12]. It is often fragmented, dispersed, and rarely standardized [12,13,15-21]. It is no secret that the EHRs do not share well across organizational lines, but with unstructured data, even within the same organization, unstructured data is difficult to aggregate and analyze. It is no wonder that 61% of

XSL•FO
RenderX

the articles analyzed listed this as a concern; big data analytics will need to address this large challenge.

Research data within the health care sector is more heterogeneous than the research data produced within other research fields [3,5,12]. Data from both research and public health is often produced in large volumes [15,22,23]. Another structure-related issue results from the changing health care fee-for-service care model [4]. Finally, big data will need to address issues with the transparency of metadata [16,24].

Security Issues

There are considerable privacy concerns regarding the use of big data analytics, specifically in health care given the enactment of Health Insurance Portability and Accountability Act (HIPAA) legislation [15]. Data that is made available on open source is freely available and, hence, highly vulnerable [12,13,18,20]. Further, due to the sensitivity of health care data, there are significant concerns related to confidentiality [25,26]. Moreover, this information is centralized, and as such, it is highly vulnerable to attacks [25]. For these reasons, enabling privacy and security is very important, as illustrated by a frequency of mention in 50% of the literature reviewed.

Data Standardization Issues

Although the EHRs share data within the same organization, intra-organizational, EHR platforms are fragmented, at best. Data is stored in formats that are not compatible with all applications and technologies [13,22]. This lack of data standardization also causes problems in transfer of that data [5,25]. It complicates data acquisition and cleansing [5,25,26]. About 39% of the literature mentioned this challenge.

Limited interoperability poses a large challenge for big data, as data is rarely standardized [12,13,16,22]. This leaves big data to face issues related to the acquisition and cleansing of data into a standardized format to enable analysis and global sharing [13,17,23,25,27]. With globalization of data, big data will have to deal with a variety of standards, barriers of language, and different terminologies.

Storage and Transfers

Data generation is inexpensive compared with the storage and transfer of the same. Once data is generated, the costs associated with securing and storing them remain high [25]. Costs are also incurred with transferring data from one place to another as well as analyzing it [14,21,22]. Some researchers have been able to combine the themes of Data structure and Storage and transfers when they illustrate how structured data can be easily stored, queried, analyzed, and so forth, but unstructured data is not as easily manipulated [13]. Cloud-based health information technology has the additional layer of security associated with the extraction, transformation, and loading of patient-related data [27]. The use of big data should address issues related to increased expenditures as well as the transmittance of secure or insecure information. About 28% of the literature mentioned this challenge.

Managerial Issues

Data governance will need to move up on the priority list of organizations, and it should be treated as a primary asset instead of a by-product of the business [15]. Data ownership and data stewardship should create new roles in business that consider big data analytics [15], and new partnerships will need to be brokered when sharing data [23,24,27]. About 14% of the literature mentioned this point.

Lack of Appropriate Skills

It is important that health care workers are also kept up to date with the use of constantly changing technology, techniques, and a constantly moving standard of care [5,24]. Due to the constant evolution of technology, there exist populations of individuals lacking specific skills; as such this is also a significant continuing barrier to the implementation of big data [12]. About 11% of the literature expressed this challenge.

Inaccuracies (Veracity)

Self-reported data is extensively used in health care, and so it is crucial that the data collected in this manner be consistent [12]. Keeping information current as well as accurate is another challenge of data collection. Precision of data is also needed to provide accurate information [12]. Only 4% of the literature mentioned this challenge.

Regulatory Compliance Issues

Health care organizations should be aware of the various legal issues that can surface in the process of managing high volume of sensitive information. Organizations implementing big data analytics as a part of their information systems will have to comply with a significant amount of standards and regulatory compliance issues specific to health care [28]. Only 4% of the literature mentioned this challenge.

Real-Time Analytics (Velocity)

One of the key requirements in health care is to be able to utilize big data in real time. Real time is defined by enabling the use of applications such as cloud computing to view said data in real time. The use of these technologies leads to issues of security and privacy within patient information [12]. Only 4% of the literature mentioned this challenge. Challenges most often mentioned or discussed were data structure (17/28, 61%), security (14/28, 50%), data standardization (11/28, 39%), and data storage and transfers (8/28, 29%). The other five challenges comprised less than 15% of the observations.

Opportunities for Big Data in Health Care

Fourteen themes emerged under the category of opportunities: improve quality of care, managing population health, early detection of diseases, data quality, structure, and accessibility, improve decision making, cost reduction, patient-centric care, enhances personalized medicine, globalization, fraud detection, and health-threat detection. Examples of each theme are listed in Table 2. A total of 113 observations were made for opportunities.

XSL•FO
RenderX

Table 2. Themes that emerged from the opportunities for big data in health care.

Themes	Examples	Number of articles (n)	Articles themes appeared in	% of total articles (N=28)
Improve quality of care	Improve efficiency Improve outcomes Reduce waste Reduce readmissions Increased productivity and performance Risk reduction Process optimization	18	2, 4, 5, 6, 8-13, 18-20, 22-25, 27	64%
Managing population health	Managing population health	17	2, 5, 8-10, 12-14, 16, 18-20, 23, 25, 26, 28	61%
Early detection of diseases	Predicting epidemics Disease monitoring Health tracking Adopt and track healthier behaviors Predicting patient vulnerability Improved treatments	17	2, 4, 5, 7-13, 15, 18-20, 23, 24, 28	61%
Data quality, structure, and accessibility	Large volumes Wide variety Creating transparency High-velocity capture Access to primary data Reusable data Weed out unwanted data Open source—free access	16	2, 4, 6, 9, 11, 12, 16, 18, 20- 23, 25-28	57%
Improve decision making	Evidence-based medicine New treatment guidelines Accuracy in information	11	2,-4, 7, 9, 12, 16, 20, 22, 23, 24	39%
Cost reduction	Inexpensive Reducing health care spending	10	1, 3, 4, 7, 9, 11, 12, 14, 16, 18	36%
Patient-centric health care	Empowering patients Patients making informed decisions Increased communication	8	2, 3, 5, 12, 14, 20, 22, 24	29%
Enhancing personalized medicine	Targeted approach	6	4-6, 24, 25, 28	24%
Globalization	Widely accessible Global sharing Leveraging knowledge and practices Knowledge dissemination	6	2, 6-8, 10, 20	24%
Fraud detection	Fraud detection	3	8, 12, 28	11%
Health-threat detection	Health-threat detection	1	7	4%

Despite the challenges that big data needs to overcome, the advanced analytics that are promised through big data offer tremendous opportunities for most stakeholders in the health care industry (patient, provider, and payer). More than 64% of the articles analyzed focused on quality improvement and more than 60% on managing population health and early detection

XSL•FO
RenderX

of diseases through big data analytics. If even some of the opportunities of big data are realized, they can radically change patient outcomes and the way decisions are made by providers, and help solve some macro-level issues related to health care within countries such as the United States (cost, quality, and access).

Improve Quality of Care

Big data has the potential and ability to improve the quality and efficiency of care [5,15,23,29-31]. Big data offers an ability to predict outcomes using the available primary or historical data and provide proof of benefit that could change established, industry-wide standards of care [25,28]. Leveraging technology at the patient end can also help with medication adherence [23,25]. This will most certainly play an important role in improving outcomes [2,13] and improve the health-related quality of life [20,26,32].

Quality of care will also be improved by reducing waste of information, which will reduce inefficiencies [13,26]. This will also assist in analyzing real-time resource utilization productivity [13]. Quality can also be improved by reducing the rates of readmissions, increasing operational efficiencies, and improving performance [5,12,13]. About 64% of the literature mentioned this opportunity.

Managing Population Health

The management of population health and the early detection of diseases were topics that the authors thought would have highly similar results after the analysis. Although there was a large overlap between the 2 themes, there was also specific variation between them. So, the researchers chose to keep them separate. The theme of managing population health focused on special populations rather than public health.

Big data analytics define populations at a finer level of granularity than has ever been previously achieved [5,14,15,33]. It can help in managing the overall health of a population as well as specific individual health [13,26,29]. Big data can enable population health management from a local or global perspective [31,34]. This capability becomes more salient from the global perspective when considering the aging of the population and age-related health issues shared by many populations and subpopulations, many of which are underserved [17,19,21,24,28,32]. About 61% of the literature mentioned this opportunity.

Early Detection of Diseases

Big data allows for the early detection of diseases, which aids in clinical objectives related to achieving improved treatments and higher patient outcomes [12,13,15,22,25]. It is in this area that the authors found great promise in age-related illness and disease. Along with early detection, big data analytics can also help in the prevention of a wide range of deadly illnesses and personalized disease management and monitoring [5,19,21,22,29,34]. It enables providers to track healthy behaviors and helps patients in monitoring their respective conditions [25,32,33]. This capability holds great potential when faced with either age-related diseases, or worldwide health issues such as cardiology [16,22,28,31,34]. About 61% of the literature mentioned this opportunity.

Data Quality, Structure, and Accessibility

Literature suggests that big data enables rapid capture of data and the conversion of primary, raw and unstructured data into meaningful information [15,17,31,34]. New knowledge can then be generated from high volumes of effective data, enabling reuse of the data [15,20,21,32,33]. Open-source technology increases accessibility to and transparency of the data [12,25,26,30,35]. Finally, data quality can be maintained using analytics to get rid of unnecessary information [27]. About 57% of the literature mentioned this opportunity.

Improve Decision Making

Big data enables appropriate use of evidence-based medicine and helps health care providers make more informed decisions [12,13,15,22]. This, in turn, improves the quality of care provided to the patients [16,31,36]. Remote monitoring, patient profile analytics, and genomic analytics are examples of other applications that influence the decision-making process [13,25].

Decision-making process can be highly optimized by the availability of accurate and up-to-date information, as decision making is influenced by the generation of new practices and treatment guidelines within clinical research. Allowing big data to influence decision making will allow for a faster and simpler process. This is done by either supporting or replacing human decision making. About 39% of the literature mentioned this opportunity.

Cost Reduction

The literature suggests that the decrease in cost of the elements of computing, such as storage and processing, leads to a decrease in the cost of data-intensive tasks [2,13]. This pass-through of savings will be seen across the spectrum of medicine [24,36] and the health care workforce [25]. Savings will be realized through more cost-effective treatments and monitoring to improve medication adherence [25,31] and through the reduction of costly transportation costs, as is experienced in cardiology [12,17,22,34]. About 36% of the literature mentioned this opportunity.

Patient-Centric Care

Increasing the use of technology is slowly changing the direction of the health care sector from disease-centric care toward patient-centric care [5]. Big data will play a significant role in this transformation [37]. It will allow the information to be delivered to patients directly and empower them to play an active part in their care [5,15,27]. When patients are provided with the appropriate information, it will influence their decision making and allow them to make informed decisions [13,24]. Informed decisions will also be influenced by increased communication between patients, providers, as well as their communities [5,24,32,36]. About 29% of the literature mentioned this opportunity.

Enhancing Personalized Medicine

With the use of big data, the objectives of personalized medicine can be translated into clinical practice [5,25,30]. Access to and processing of large volumes of data should enable a personalized patient-specific record of risks of disease [25,29,32]. Big data

XSL•FO
RenderX

applications aim to make this process more efficient [12]. About 24% of the literature mentioned this opportunity.

Globalization

Big data will actively help in disseminating the knowledge acquired from the data collected [15,22,30]. Big data plays an active role in leveraging the practices and knowledge not only regionally but globally [12,15,29]. By globalizing data, it is made more widely accessible and providers may access new information from all regions [22,23,32]. About 24% of the literature mentioned this opportunity.

Fraud Detection

One of the most significant benefits offered by big data is that it is instrumental in detecting fraud in an efficient and effective manner [13,23]. For example, the unauthorized use of specific user accounts by third parties can be minimized [21]. Only about 11% of the literature mentioned this opportunity.

Health-Threat Detection

Big data offers opportunity for improving capabilities of threat detection quickly and more accurately. This can be especially beneficial for government use [22]. Big data augments the current acquisition of protection against the increasing threats of foreign countries, criminals, terrorists, and others. Only 3.6% of the literature mentioned this opportunity.

Opportunities most often mentioned or discussed were improve quality of care (18/28, 64%), managing population health (17/28, 61%), early detection of diseases (17/28, 60.7%), data quality structure and accessibility (16/28, 57%), improve decision making (11/28, 39.3%), cost reductions (10/28, 36%), patient-centric health care (8/28, 29%), enhancing personalized medicine (6/28, 24%), and globalization (6/28, 24%). The other two opportunities each comprised less than 15% of the observations.

Discussion

Summary of Evidence

Although the integration of big data is well underway in industries such as finance and advertising, it has not yet fully assimilated into health care. Challenges and opportunities were made quite clear in the articles analyzed in this review. Three of the 4 Vs (volume, velocity, and variety) were consistently adhered to. The fourth V, veracity, was found, but rarely listed by name. Tables 1 and 2 provide insightful information that is previously unpublished. These tables identify challenges and opportunities and illustrate their frequency of mention in the literature. This information is helpful to other researchers and innovators because it provides direction and proper emphasis of research effort. The listed challenges and opportunities are ordered by their frequency found in the literature.

Limitations

A big limitation in this review is the low number of articles used in the analysis. If we were to do this over again, we would query another database to see whether additional articles were available for analysis.

Selection bias seems to exist in any study. Our control for selection bias was the initial research up front to agree on a definitive definition of the concept of big data, and our consensus meetings to discuss findings. The consensus meetings offered great value to the process because they enabled the group to hear the focus of an individual and either provide feedback to confirm the focus or agree that the unique focus was warranted for all the articles in the review.

Another bias that we discuss regularly is publication bias. Journals tend to publish results that are statistically significant, which inherently limits the publication of research that may not reach that level. Our control for publication bias was to include Google Scholar in our search. Our intent was to identify material in lesser-known journals that might not be indexed in PubMed (MEDLINE) or CINAHL.

Conclusions

Big data and the use of advanced analytics have the potential to advance the way in which providers leverage technology to make informed clinical decisions. However, the vast amounts of information generated annually within health care must be organized and compartmentalized to enable universal accessibility and transparency between health care organizations.

Our systematic literature review revealed both challenges and opportunities that big data offers to the health care industry. The literature mentioned the challenges of data structure and security in at least 50% of the articles reviewed. The literature also mentioned the opportunities of increased quality, better management of population health, early detection of disease, and data quality structure and accessibility in at least 50% of the articles reviewed. These findings identify foci for future research.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Summary or relevance of cited work.

[PDF File (Adobe PDF File), 33KB - medinform_v4i4e38_app1.pdf]

References

XSL•FO
RenderX

1. PRISMA. Welcome to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) website! 2009. URL: http://www.prisma-statement.org/ [accessed 2015-07-30] [WebCite Cache ID 6aPmhFUMM]
2. McAfee A, Brynjolfsson E. Big data: the management revolution. Harv Bus Rev 2012 Oct;90(10):60-6, 68, 128. [Medline: 23074865]
3. Heudecker N. Hype Cycle for Big Data. Gartner. 2013 Jul 31. URL: https://www.gartner.com/doc/2574616/hype-cycle-big-data- [accessed 2016-11-08] [WebCite Cache ID 6lsI6Sxxr]
4. Kayyali B, Knott D, Van Kuiken S. The big-data revolution in US health care: accelerating value and innovation. McKinsey & Company. 2013 Apr. URL: https://digitalstrategy.nl/wp-content/uploads/E2-2013.04-The-big-data-revolution-in-US-health-care-Accelerating-value-and-innovation.pdf [accessed 2016-11-11] [WebCite Cache ID 6lvquouez]
5. Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. J Gen Intern Med 2013 Sep;28(Suppl 3):S660-S665 [FREE Full text] [doi: 10.1007/s11606-013-2455-8] [Medline: 23797912]
6. US Department of Health and Human Services. Genomics Data Sharing. 2014. URL: https://gds.nih.gov/17summary_dbGaP_statistics.html [WebCite Cache ID 6m62Idais]
7. Feldman B, Martin E, Skotnes T. Big data in healthcare hype and hope. GHDonline. 2012 Oct. URL: https://www.ghdonline.org/uploads/big-data-in-healthcare_B_Kaplan_2012.pdf [accessed 2016-11-09] [WebCite Cache ID 6lt9sqaYc]
8. 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]
9. US Department of Health and Human Services. 28. Medicare and Medicaid Programs; Electronic Health Record Incentive Program. US Government Printing Office. Federal Register. 2010 Jul 28. URL: https://www.federalregister.gov/documents/2010/07/28/2010-17207/medicare-and-medicaid-programs-electronic-health-record-incentive-program [accessed 2016-11-09] [WebCite Cache ID 6ltJ4cdh3]
10. US Department of Health and Human Services. State Health Information Exchange Cooperative Agreement Program. US Printing Office. HealthIT. 2011. URL: https://www.healthit.gov/policy-researchers-implementers/state-health-information-exchange [accessed 2016-11-09] [WebCite Cache ID 6ltJbJxRk]
11. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013 Apr 3;309(13):1351-1352. [doi: 10.1001/jama.2013.393] [Medline: 23549579]
12. Jee K, Kim GH. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. Healthc Inform Res 2013 Jun;19(2):79-85 [FREE Full text] [doi: 10.4258/hir.2013.19.2.79] [Medline: 23882412]
13. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014;2:3 [FREE Full text] [doi: 10.1186/2047-2501-2-3] [Medline: 25825667]
14. Song TM, Song J, An JY, Hayman LL, Woo JM. Psychological and social factors affecting Internet searches on suicide in Korea: a big data analysis of Google search trends. Yonsei Med J 2014 Jan;55(1):254-263 [FREE Full text] [doi: 10.3349/ymj.2014.55.1.254] [Medline: 24339315]
15. Fernandes L, O'Connor M, Weaver V. Big data, bigger outcomes: healthcare is embracing the big data movement, hoping to revolutionize HIM by distilling vast collection of data for specific analysis. J AHIMA 2012 Oct;83(10):38-43; quiz 44. [Medline: 23061351]
16. Kim T, Park K, Yi S. A Big Data Framework for u-Healthcare Systems Utilizing Vital Signs. Presented at: Computer, Consumer and Control (IS3C), 2014 International Symposium; 10-12 June, 2014; Taichung, Taiwan. IEEE; 2014 Jun 30. [doi: 10.1109/IS3C.2014.135]
17. Augustine D. Leveraging big data analytics and Hadoop in developing India's healthcare service. Int J Comput Applications 2014 Mar;89(16):44-50 [FREE Full text]
18. Jiang P, Winkley J, Zhao C, Munnoch R, Min G, Yang LT. An intelligent information forwarder for healthcare big data systems with distributed wearable sensors. IEEE Systems Journal 2016 Sep;10(3):1147-1159. [doi: 10.1109/JSYST.2014.2308324]
19. Hrovat G, Stiglic G, Kokol P, Ojsteršek M. Contrasting temporal trend discovery for large healthcare databases. Comput Methods Programs Biomed 2014;113(1):251-257. [doi: 10.1016/j.cmpb.2013.09.005] [Medline: 24120407]
20. Baro E, Degoul S, Beuscart R, Chazard E. Toward a literature-driven definition of big data in healthcare. Biomed Res Int 2015;2015:639021 [FREE Full text] [doi: 10.1155/2015/639021] [Medline: 26137488]
21. Naqishbandi T, Imthyaz Sheriff C, Qazi S. Big data, CEP and IoT: redefining holistic healthcare information systems and analytics. Int J Eng Res and Technol 2015;4(1):1-6 [FREE Full text]
22. Hsieh JC, Li AH, Yang CC. Mobile, cloud, and big data computing: contributions, challenges, and new directions in telecardiology. Int J Environ Res Public Health 2013 Nov 13;10(11):6131-6153 [FREE Full text] [doi: 10.3390/ijerph10116131] [Medline: 24232290]
23. Sepulveda MJ. From worker health to citizen health: moving upstream. J Occup Environ Med 2013 Dec;55(12 Suppl):S52-S57 [FREE Full text] [doi: 10.1097/JOM.0000000000000033] [Medline: 24284749]

XSL•FO
RenderX

24. Baker TB, Gustafson DH, Shah D. How can research keep up with eHealth? Ten strategies for increasing the timeliness and usefulness of eHealth research. J Med Internet Res 2014;16(2):e36 [FREE Full text] [doi: 10.2196/jmir.2925] [Medline: 24554442]
25. Mohr DC, Burns MN, Schueller SM, Clarke G, Klinkman M. Behavioral intervention technologies: evidence review and recommendations for future research in mental health. Gen Hosp Psychiatry 2013;35(4):332-338 [FREE Full text] [doi: 10.1016/j.genhosppsych.2013.03.008] [Medline: 23664503]
26. Mancini M. Exploiting big data for improving healthcare services. J e-Learning Knowledge Soc 2014;10(2):1-11 [FREE Full text]
27. Youssef AE. A framework for secure healthcare systems based on big data analytics in mobile cloud computing environments. Int J Ambient Syst Appl 2014;2(2):1-11. [doi: 10.5121/ijasa.2014.2201]
28. Schilsky RL, Michels DL, Kearbey AH, Yu PP, Hudis CA. Building a rapid learning health care system for oncology: the regulatory framework of CancerLinQ. J Clin Oncol 2014 Aug 1;32(22):2373-2379 [FREE Full text] [doi: 10.1200/JCO.2014.56.2124] [Medline: 24912897]
29. Moore P, Thomas A, Tadros G, Xhafa F, Barolli L. Detection of the onset of agitation in patients with dementia: real-time monitoring and the application of big-data solutions. IJSSC 2013;3(3):136-154. [doi: 10.1504/IJSSC.2013.056405]
30. Wang P, Chen Z. Traditional Chinese medicine ZHENG and OMICS convergence: a systems approach to post-genomics medicine in a global world. OMICS 2013 Sep;17(9):451-459. [doi: 10.1089/omi.2012.0057] [Medline: 23837436]
31. Lamarche-Vadel A, Pavillon G, Aouba A, Johansson LA, Meyer L, Jougla E, et al. Automated comparison of last hospital main diagnosis and underlying cause of death ICD10 codes, France, 2008-2009. BMC Med Inform Decis Mak 2014;14:44 [FREE Full text] [doi: 10.1186/1472-6947-14-44] [Medline: 24898538]
32. Howren MB, Vander Weg MW, Wolinsky FD. Computerized cognitive training interventions to improve neuropsychological outcomes: evidence and future directions. J Comp Eff Res 2014 Mar;3(2):145-154. [doi: 10.2217/cer.14.6] [Medline: 24645688]
33. Wlodarczyk TW, Hacker TJ. Current trends in predictive analytics of big data. Int J Big Data Intel 2014;1(3):172-122. [doi: 10.1504/IJBDI.2014.066326]
34. Sengupta PP. Intelligent platforms for disease assessment: novel approaches in functional echocardiography. JACC Cardiovasc Imaging 2013 Nov;6(11):1206-1211 [FREE Full text] [doi: 10.1016/j.jcmg.2013.09.003] [Medline: 24229773]
35. Issa NT, Byers SW, Dakshanamurthy S. Big data: the next frontier for innovation in therapeutics and healthcare. Expert Rev Clin Pharmacol 2014 May;7(3):293-298 [FREE Full text] [doi: 10.1586/17512433.2014.905201] [Medline: 24702684]
36. Beveridge R, Fox J, Higgins SA, Kohn M, Mahoney JJ, Newcomer LN, et al. Roundtable—the changing oncology landscape: evolution or revolution? J Natl Compr Canc Netw 2013 May;11(5 Suppl):636-638. [Medline: 23704232]
37. Kaushik K, Kapoor D, Varadharajan V, Nallusamy R. Disease management: clustering-based disease prediction. Int J Collaborative Enterprise 2014;4(1-2):69-82. [doi: 10.1504/IJCENT.2014.065047]

Abbreviations

ARRA: American Recover and Reinvestment Act
EHR: electronic health record
HIE: Health Information Exchange
HIPAA: Health Insurance Portability and Accountability Act
HITECH: Health Information Technology for Economic and Clinical Health
MeSH: Medical Subject Headings
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-analysis

Edited by G Eysenbach; submitted 19.11.15; peer-reviewed by J Bian, D Maslove, MA Mayer, S Seevanayanagam, L Toldo; comments to author 03.01.16; revised version received 27.07.16; accepted 28.09.16; published 21.11.16.

Please cite as:
Kruse CS, Goswamy R, Raval Y, Marawi S
Challenges and Opportunities of Big Data in Health Care: A Systematic Review
JMIR Med Inform 2016;4(4):e38
URL: http://medinform.jmir.org/2016/4/e38/
doi:10.2196/medinform.5359
PMID:27872036

©Clemens Scott Kruse, Rishi Goswamy, Yesha Raval, Sarah Marawi. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 21.11.2016. This is an open-access article distributed under the terms of the Creative Commons

Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Increasing Complexity in Rule-Based Clinical Decision Support: The Symptom Assessment and Management Intervention

David F Lobach[1,2], MD, PhD, MS, FACMI; Ellis B Johns[3,4], MD, MS; Barbara Halpenny[5], MA; Toni-Ann Saunders[5], MPH; Jane Brzozowski[6], MS; Guilherme Del Fiol[7], MD, PhD; Donna L Berry[5], PhD, RN, FAAN; Ilana M Braun[8], MD; Kathleen Finn[9], RN, MSN, AOCN, NP; Joanne Wolfe[8], MD; Janet L Abrahm[8], MD; Mary E Cooley[5], PhD, CRNP, FAAN

[1]School of Medicine, Department of Community & Family Medicine, Duke University, Durham, NC, United States

[2]Klesis Healthcare, Durham, NC, United States

[3]Family Medicine of Albemarle, Charlottesville, VA, United States

[4]Medengineers Informatics, Charlottesville, VA, United States

[5]Dana-Farber Cancer Institute, The Phyllis F. Cantor Center, Boston, MA, United States

[6]Independent Clinical Informatics Consultant, Boston, MA, United States

[7]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, United States

[8]Department of Psychosocial Oncology and Palliative Care, Dana-Farber Cancer Institute, Boston, MA, United States

[9]City of Hope, Clinical Trials Office, Duarte, CA, United States

Corresponding Author:
David F Lobach, MD, PhD, MS, FACMI
School of Medicine
Department of Community & Family Medicine
Duke University Medical Center
Duke University
Box 3886
Durham, NC, 22710
United States
Phone: 1 919 438 2346
Fax: 1 919 681 7085
Email: David.Lobach@klesishealthcare.com

Abstract

Background: Management of uncontrolled symptoms is an important component of quality cancer care. Clinical guidelines are available for optimal symptom management, but are not often integrated into the front lines of care. The use of clinical decision support (CDS) at the point-of-care is an innovative way to incorporate guideline-based symptom management into routine cancer care.

Objective: The objective of this study was to develop and evaluate a rule-based CDS system to enable management of multiple symptoms in lung cancer patients at the point-of-care.

Methods: This study was conducted in three phases involving a formative evaluation, a system evaluation, and a contextual evaluation of clinical use. In Phase 1, we conducted iterative usability testing of user interface prototypes with patients and health care providers (HCPs) in two thoracic oncology clinics. In Phase 2, we programmed complex algorithms derived from clinical practice guidelines into a rules engine that used Web services to communicate with the end-user application. Unit testing of algorithms was conducted using a stack-traversal tree-spanning methodology to identify all possible permutations of pathways through each algorithm, to validate accuracy. In Phase 3, we evaluated clinical use of the system among patients and HCPs in the two clinics via observations, structured interviews, and questionnaires.

Results: In Phase 1, 13 patients and 5 HCPs engaged in two rounds of formative testing, and suggested improvements leading to revisions until overall usability scores met a priori benchmarks. In Phase 2, symptom management algorithms contained between 29 and 1425 decision nodes, resulting in 19 to 3194 unique pathways per algorithm. Unit testing required 240 person-hours, and integration testing required 40 person-hours. In Phase 3, both patients and HCPs found the system usable and acceptable, and offered suggestions for improvements.

Conclusions: A rule-based CDS system for complex symptom management was systematically developed and tested. The complexity of the algorithms required extensive development and innovative testing. The Web service-based approach allowed remote access to CDS knowledge, and could enable scaling and sharing of this knowledge to accelerate availability, and reduce duplication of effort. Patients and HCPs found the system to be usable and useful.

(JMIR Med Inform 2016;4(4):e36) doi:10.2196/medinform.5728

KEYWORDS

rule-based clinical decision support; clinical algorithms; Web services; software as a service; symptom management; patient-reported outcomes; lung cancer

Introduction

Clinical decision support (CDS) derived from clinical algorithms (ie, rule-based) is essential for improving the quality and safety of health care [1]. In spite of the critical nature of this resource, much of rule-based CDS to date has been relatively simplistic, and few examples of complex decision algorithms with dozens of decision points have been implemented [2,3]. As increasingly complex clinical protocols are implemented through CDS, innovative approaches will be required to thoroughly and rigorously validate the accuracy of these CDS systems [4].

In order to fulfill the clinical expectations of CDS in the future, the next generation of rule-based CDS will need to mature to: (1) accommodate increasing clinical complexity; (2) respond to current patient status by incorporating real-time clinical information, including patient-reported data; and (3) increase efficiency by allowing for scaling and portability through reuse of decision logic by separating the end user application from the decision engine. In this project, we developed a CDS system that supported all three of these features. This system supported the complex challenge of simultaneously managing multiple symptoms (anxiety, depression, dyspnea, fatigue, and pain) in patients with lung cancer, the collection of real-time symptom data from patients, and potential reutilization of algorithm knowledge via Web services.

Symptom management in lung cancer patients is complex, and uncontrolled symptoms have been associated with increased emotional distress, decreased health-related quality of life, and even decreased survival [5-9]. The majority of lung cancer patients have high levels of disease-related symptomatology, as well as psychological distress at presentation [10-14]. Optimal management requires attention to multiple symptoms. To date, the majority of studies aiming to enhance symptom management have addressed the treatment of individual symptoms [15-19]. New approaches to manage multiple distressing symptoms are needed. National groups have called for improving symptom management and palliative care across the cancer continuum, and for supporting improved quality of care with the use of health care information technology [20-22]. In a prior project, we convened multidisciplinary panels of clinical experts to develop computable symptom management algorithms for multiple symptoms based on national clinical practice guidelines [23]. These algorithms provided recommendations for specific pharmacological and behavioral interventions–tailored to a patient’s age, comorbidities, laboratory values, current medications, and patient-reported symptom severity–to manage anxiety, depression, dyspnea, fatigue, and pain. The complex algorithms, and their integration with one another, approximated the cognitive processes of clinical experts and considered multiple factors that may aggravate and/or alleviate common cancer symptoms. Further information regarding the expert panel and processes used to develop the computable algorithms has been published previously [23].

In this paper we report on the development, testing, and contextual evaluation of the Symptom Assessment and Management Intervention for Lung cancer (SAMI-L) CDS system that was based on these algorithms, in two hospital-based clinics. In *Phase 1*, our objective was to develop usable and acceptable user interfaces to accurately capture the patient-reported and clinical data required to process the algorithms, and to display guideline-based recommendations in interpretable and actionable ways to health care providers (HCPs). In *Phase 2*, our objective was to program and test the accuracy of the algorithms and the integrated system. In *Phase 3*, our objective was to evaluate the use of the system by patients and HCPs in the clinical setting.

System Description

The SAMI-L system consists of three components: (1) a Web-based assessment tool for collecting patient-reported data on symptom severity, medications, and laboratory values using a touch screen notebook computer. This tool uses standardized patient-reported outcome questionnaires that have been used previously with cancer patients, and are among the most commonly used measures in such studies [24-27]; (2) a *decision engine* known as the System for Evidence-Based Advice through Simultaneous Transaction with an Intelligent Agent Across a Network (SEBASTIAN) [28], accessed remotely using Web services; and (3) printed reports for clinicians that summarize patient data and present patient-specific recommendations (Figure 1).

Figure 1 identifies the components of the SAMI-L system and the data flow between these components. Patients and research assistants entered data on a touch screen notebook computer in the clinic waiting area. These data were then transmitted, with a session identification number and no personal health information, through the PROQuest server to the SEBASTIAN decision support engine using Web services. After processing the data, the recommendations were returned from the decision engine through Web services to the PROQuest server where they were formatted into patient reports. These reports were then printed and delivered to the healthcare provider in the examination room.

XSL•FO
RenderX

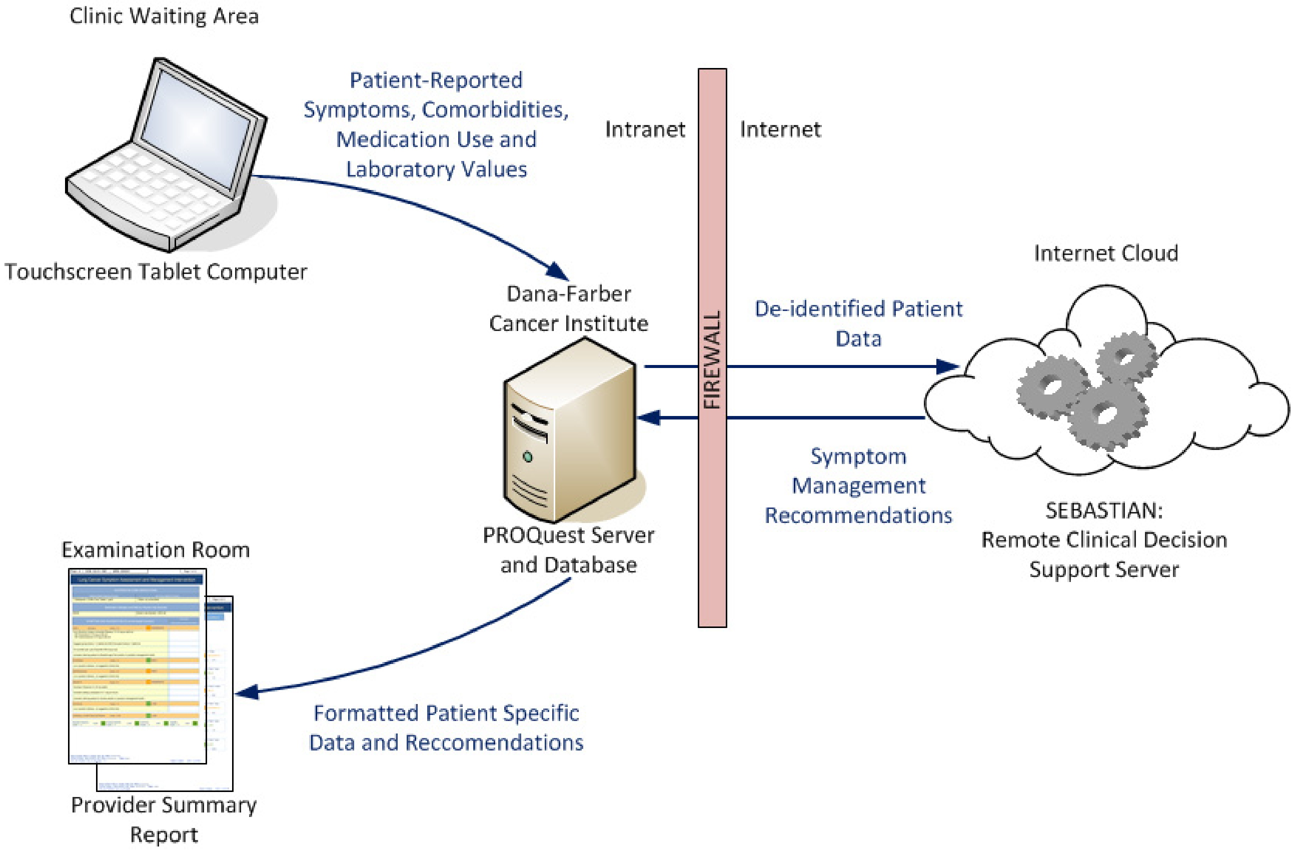
The *decision engine* was built using a Web service-based CDS tool known as SEBASTIAN [28]. The SEBASTIAN system is one of the initial decision engines that implemented CDS using Web services [29]. This system provided the foundation for the evolving HL7 Decision Support Service standard and has been described previously [28].

SEBASTIAN receives data from remote client applications structured in a common *language* known as eXtensible Markup Language (XML). Using this Web service framework, decision logic can be centralized in SEBASTIAN for use by many systems at different sites, thus enabling the sharing of computable knowledge across multiple remote locations [30]. The complex symptom management algorithms were represented in the form of procedural rules, and implemented into SEBASTIAN using an object-oriented computer programming language (Java). In order to generate specific symptom management care recommendations, the symptoms, medications, and laboratory values were submitted as Web service requests from a server in Boston, Massachusetts to a cloud-based server for processing by the SEBASTIAN inference engine [31]. Submitted patient information was distinguished by a unique session identifier so that only nonidentifying patient information was transmitted to the CDS server. Complete traversal of each decision node was critical for generating correct recommendations, so we programmed the SAMI-L system to function only if all required data were available. Accordingly, each clinical rule would determine that all of the required data were present before running. If data were missing, the system would send a message stating that the available data were insufficient to run the algorithm.

SAMI-L also generated a printed report for clinicians to use during the clinical visit (Figure 2). This two-page report included a summary of patient-reported data along with patient-specific recommendations based on the symptom management algorithms. This information was presented in lists, tables, charts with color coding, and trend graphs to make it easily consumable by clinicians. The symptom management guidance was based on the severity of a patient's symptoms. Guidance included specific suggestions for use of medications (including recommendations to initiate medications or explicit adjustments for medication doses), laboratory tests, supportive care referrals (ie, social work, palliative care, psychiatry), and use of a self-care symptom management toolkit for patients that provided behavioral self-care suggestions [32,33].

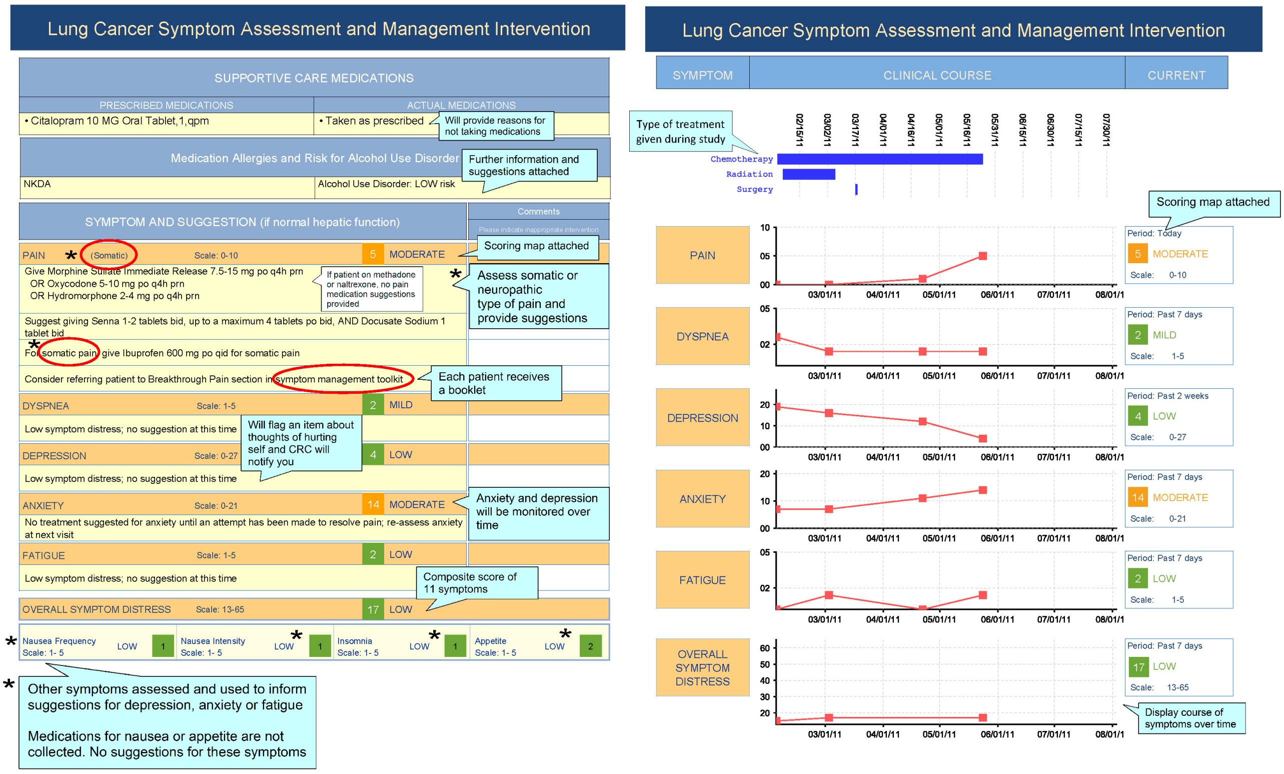
The left panel of Figure 2 provides the data from which the care recommendations were derived, including the current medications, medication allergies, alcohol use history, and the patient-reported level of distress by individual symptoms. Level of patient symptom distress was color-coded with green, yellow, and red to indicate increasing levels of distress. Explicit, patient-specific care guidance recommendations are provided with each individual symptom, as determined from the care algorithm. The right panel of Figure 2 shows a time course summary of a patient's treatments, and a cumulative graphical summary of changes in a patient's levels of symptom distress over time by each individual symptom. Figure 2 first appeared in Cooley et al [34].

Figure 1. SAMI-L system architecture and overview.


Clinic Waiting Area
Touchscreen Tablet Computer
Patient-Reported Symptoms, Comorbidities, Medication Use and Laboratory Values
Intranet
Internet
FIREWALL
Dana-Farber Cancer Institute
PROQuest Server and Database
De-identified Patient Data
Symptom Management Recommendations
Internet Cloud
SEBASTIAN: Remote Clinical Decision Support Server
Examination Room
Provider Summary Report
Formatted Patient Specific Data and Reccomendations

XSL•FO
RenderX

Figure 2. Sample report produced by SAMI-L.



Lung Cancer Symptom Assessment and Management Intervention

SUPPORTIVE CARE MEDICATIONS

PRESCRIBED MEDICATIONS	ACTUAL MEDICATIONS
• Citalopram 10 MG Oral Tablet,1,qpm	• Taken as prescribed

Will provide reasons for not taking medications

Medication Allergies and Risk for Alcohol Use Disorder

NKDA	Alcohol Use Disorder: LOW risk

Further information and suggestions attached

SYMPTOM AND SUGGESTION (if normal hepatic function)

Comments

Please indicate inappropriate intervention

PAIN * (Somatic) Scale: 0-10	5	MODERATE	
Give Morphine Sulfate Immediate Release 7.5-15 mg po q4h prn OR Oxycodone 5-10 mg po q4h prn OR Hydromorphone 2-4 mg po q4h prn			
Suggest giving Senna 1-2 tablets bid, up to a maximum 4 tablets po bid, AND Docusate Sodium 1 tablet bid			
* For somatic pain give Ibuprofen 600 mg po qid for somatic pain			
Consider referring patient to Breakthrough Pain section in symptom management toolkit			
DYSPNEA Scale: 1-5	2	MILD	
Low symptom distress; no suggestion at this time			
DEPRESSION Scale: 0-27	4	LOW	
Low symptom distress; no suggestion at this time			
ANXIETY Scale: 0-21	14	MODERATE	
No treatment suggested for anxiety until an attempt has been made to resolve pain; re-assess anxiety at next visit			
FATIGUE Scale: 1-5	2	LOW	
Low symptom distress; no suggestion at this time			
OVERALL SYMPTOM DISTRESS Scale: 13-65	17	LOW	

* Nausea Frequency Scale: 1- 5	LOW	1	Nausea Intensity Scale: 1- 5	* LOW	1	Insomnia Scale: 1- 5	* LOW	1	Appetite Scale: 1- 5	* LOW	2

Scoring map attached

If patient on methadone or naltrexone, no pain medication suggestions provided

* Assess somatic or neuropathic type of pain and provide suggestions

Each patient receives a booklet

Will flag an item about thoughts of hurting self and CRC will notify you

Anxiety and depression will be monitored over time

Composite score of 11 symptoms

* Other symptoms assessed and used to inform suggestions for depression, anxiety or fatigue

Medications for nausea or appetite are not collected. No suggestions for these symptoms

Lung Cancer Symptom Assessment and Management Intervention

SYMPTOM | CLINICAL COURSE | CURRENT

Type of treatment given during study

Scoring map attached

SYMPTOM	CURRENT
PAIN	Period: Today 5 MODERATE Scale: 0-10
DYSPNEA	Period: Past 7 days 2 MILD Scale: 1-5
DEPRESSION	Period: Past 2 weeks 4 LOW Scale: 0-27
ANXIETY	Period: Past 7 days 14 MODERATE Scale: 0-21
FATIGUE	Period: Past 7 days 2 LOW Scale: 1-5
OVERALL SYMPTOM DISTRESS	Period: Past 7 days 17 LOW Scale: 13-65

Display course of symptoms over time

Methods

Phase 1

In *Phase 1*, we conducted iterative usability testing of user interface prototypes with patients and HCPs in two thoracic oncology clinics. The expert panels had created computable symptom management algorithms that specified validated patient-reported symptom measures and clinical data that were required to process the algorithms in a previous project [23]. To create the patient component of the system, we constructed validated self-report symptom assessment questionnaires measuring the targeted symptoms, and data entry interfaces for required medication history and clinical variables. The questionnaires were constructed using an existing Web-based data collection platform at Dana-Farber Cancer Institute (DFCI). Patient participants were >21 years of age, English speaking, diagnosed with Stage III or IV nonsmall cell lung cancer, had limited or extensive stage small cell lung cancer or new recurrence of disease, were receiving care in the outpatient setting, and were actively receiving cancer-directed treatment. These patients were recruited for 60-minute usability test sessions. We oversampled patients from Boston Medical Center (BMC), a community-based safety-net hospital, to ensure representation of patient users with lower literacy and lower familiarity with computers. A usability interviewer from the Dana-Farber/Harvard Cancer Center Health Communication Core (HCC) used a structured interview guide to observe and elicit feedback on understanding of the assessment, ease of navigation, helpfulness of the program, the amount of time required to complete the program, and overall user satisfaction. The interviewer also observed mock reviews of medication history by the study coordinator, as would be required to obtain data needed to process the algorithms. Patient participants completed the Acceptability E-Scale [35] and a demographic questionnaire at the end of the session. Following the appropriate protocol, two or more rounds of testing were required until acceptability scores met the predefined threshold of an average score of 4 on a 5-point scale (1=low; 5=high) for each item, or a composite score of >24 across six items.

To create the HCP component of the system, prototype graphical summary reports of the CDS recommendations for symptom management were developed by a graphic designer in the HCC. We recruited eligible HCPs, who were attending physicians or nurse practitioners in the two thoracic medical oncology clinics, and randomized them to intervention or usual care arms for the trial. Participants in the intervention arm were invited to participate in formative usability testing of the reports. We conducted 30-minute usability sessions in which HCPs were presented with high fidelity mock reports of patients' current and historical symptom status, and recommended pharmacologic and behavioral interventions. A research team member followed a structured script to solicit feedback and probe understanding of layout, content, and visual style of each section of the report. Participants then completed standard usability rating questionnaires [36,37] and a demographic questionnaire. Following the appropriate protocol, two or more rounds of testing were required until acceptability scores met the predefined threshold of an average of 4 on 5-point scale (1=low; 5=high) across all items.

Phase 2

In *Phase 2*, we programmed the five complex algorithms (anxiety, depression, dyspnea, fatigue, and pain), which were derived from clinical practice guidelines, into a rules engine

XSL•FO
RenderX

that used Web services to communicate with the end-user application. We conducted unit testing of algorithms using a stack-traversal tree-spanning (STTS) methodology to identify all possible permutations of pathways through each algorithm, to validate accuracy. The symptom management algorithms defined by the expert panels required >30 unique data elements (Multimedia Appendix 1) and were developed to address multiple clinical issues for appropriate symptom management (Multimedia Appendix 2).

Multimedia Appendix 1 provides information about the data requirements that were needed to inform the algorithms to generate specific recommendations for symptom management, the standardized assessment instruments that were used to collect the data, and the source of the data collection. Multimedia Appendix 2 provides information about the type of recommendations that were provided for each of the five algorithms (anxiety, depression, dyspnea, fatigue, and pain) and the specific data elements that were required to generate those recommendations.

Algorithm Complexity

In order to quantify the complexity of the symptom management algorithms, we determined the number of decision nodes and unique pathways within each algorithm. For this purpose, we counted a *decision node* as a point within an algorithm where the logic could branch in two or more directions. In some algorithms, specific clinical parameters (ie, renal function) appeared in two or more distinct parts of the algorithm, based on when in the course of decision-making kidney function should be considered. In such cases, each instance of the renal function node would be added to the total node count for the algorithm. A *pathway* was defined as a unique sequence of branches through the algorithm that began at the entry point of the algorithm and ended at a specific end node from which no additional decision nodes followed.

After programming logic content into the SEBASTIAN decision engine, the number of decision nodes in the symptom management algorithms ranged from a low of 29 in the fatigue algorithm to a high of 1425 in the pain algorithm (Table 1). Traversal of these algorithms across all possible variable permutations identified a low of 19 unique pathways in the fatigue algorithm and a high of 3194 pathways in the pain algorithm (Table 1).

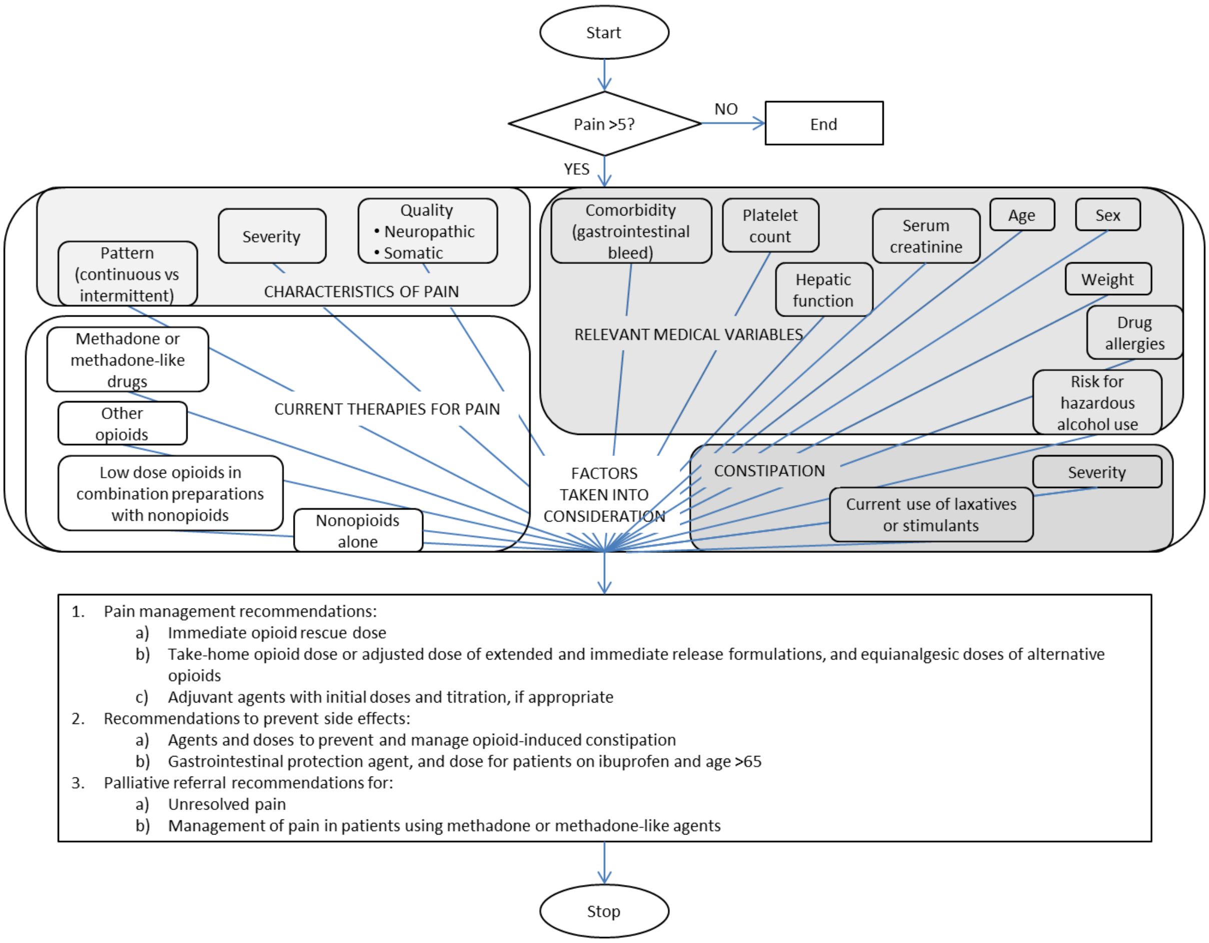
As an illustration of the complexity of the algorithm for pain management, the diagram in Figure 3 portrays the factors that were considered in generating recommendations for care guidance, along with the types of recommendations that are typically considered for patients experiencing significant levels of pain. Figure 3 displays a schematic representation of the pain algorithm to demonstrate the complexity of the logic considered for managing pain. The upper component of Figure 3 illustrates the multiple factors that were taken into consideration, in order to generate care guidance recommendations to manage pain. Factors include the characteristics of the pain, the current therapy for the pain, relevant medical variables, and issues related to opioid-induced constipation. The lower component of Figure 3 summarizes the types of recommendations that were produced, including recommendations for pain management, recommendations to prevent side effects from pain medications, and recommendations for palliative care referrals.

Table 1. Number of unique pathways and decision nodes in symptom management algorithms.

Rule	Decision Nodes	Separate Pathways
Anxiety	45	43
Depression	42	39
Fatigue	29	19
Pain	1425	3194
Dyspnea	87	113

XSL•FO
RenderX

Figure 3. Pain algorithm data components and recommendations.


Start
Pain >5?
NO
End
YES
Pattern (continuous vs intermittent)
Severity
Quality
• Neuropathic
• Somatic
CHARACTERISTICS OF PAIN
Comorbidity (gastrointestinal bleed)
Platelet count
Serum creatinine
Age
Sex
Hepatic function
Weight
Drug allergies
Risk for hazardous alcohol use
RELEVANT MEDICAL VARIABLES
Methadone or methadone-like drugs
Other opioids
Low dose opioids in combination preparations with nonopioids
Nonopioids alone
CURRENT THERAPIES FOR PAIN
FACTORS TAKEN INTO CONSIDERATION
CONSTIPATION
Severity
Current use of laxatives or stimulants
Stop

1. Pain management recommendations:
   a) Immediate opioid rescue dose
   b) Take-home opioid dose or adjusted dose of extended and immediate release formulations, and equianalgesic doses of alternative opioids
   c) Adjuvant agents with initial doses and titration, if appropriate
2. Recommendations to prevent side effects:
   a) Agents and doses to prevent and manage opioid-induced constipation
   b) Gastrointestinal protection agent, and dose for patients on ibuprofen and age >65
3. Palliative referral recommendations for:
   a) Unresolved pain
   b) Management of pain in patients using methadone or methadone-like agents

System Testing Approach

The complex care algorithms developed to address simultaneous symptom management required new methods to thoroughly and rigorously validate the accuracy of the CDS recommendations. Accordingly, a systematic approach was developed to ensure that all of the possible permutations arising from hundreds of branching pathways had been assessed. First, in order to identify errors during unit testing, the study team selected hundreds of representative instances of automatically generated test cases with predetermined recommendations. Next, the test cases were submitted to SEBASTIAN and mismatches between the newly-generated recommendations and the expected recommendations were identified. The advantage of this approach was that future changes in algorithms could be tested by running the same test cases. Results of hundreds of test cases were also manually compared to the algorithm flowcharts (approved by an expert clinician) to ensure that there were no logic errors in the algorithms.

Second, in order to identify errors during integration testing, the study team developed a set of 10 test cases. These test cases were sent to the CDS Web service from the study sites, using the data collected via SAMI-L. The recommendations generated from SEBASTIAN were reviewed by a clinical expert to ensure their accuracy. In addition, the display of patient data, and the resulting recommendations that were part of the HCP report, were verified to ensure accuracy.

Finally, we created a systematic and reusable testing approach to validate the accuracy of complex care protocols using an STTS algorithm. For each algorithm, using an XML text editor, we created an XML data input file with data parameters targeting boundary conditions for each decision node. All possible permutations for traversing all of the pathways through each protocol were created using an XML-based STTS algorithm written in Java. The data elements defining each permutation were sequentially submitted as Web service requests to the decision engine. Each resultant set of recommendations was paired with the data set used to generate the response, and Altova MapForce [38] auto-generated Java code was used to map the input and output parameters to a queryable relational database. Initially, research staff rigorously queried the database to confirm that the correct recommendations had been generated from each paired variable-input-recommendation-output data set. Based on these systematic queries, inconsistencies in the logic were identified. The development team then corrected the logical inconsistencies by modifying the flow diagram and associated algorithms to correct the erroneous logic. The testing cycle was then repeated to ensure accuracy of the decision logic. In final testing, we automated the validation of the data-recommendation pairs using a unit testing approach with a set of manually validated test cases serving as the standard (ie, if new rule input-output parameters did not correspond with input-output parameters from a validated database, then new rule logic errors would be addressed). Care recommendations

XSL•FO
RenderX

provided guidance that could directly impact patient care, so we required 100% accuracy of the generated recommendations (in terms of agreement with the stipulated algorithm) before the CDS for management of each symptom was moved to production. A clinical expert (JLA) reviewed all recommendations generated by the algorithms to ensure accuracy.

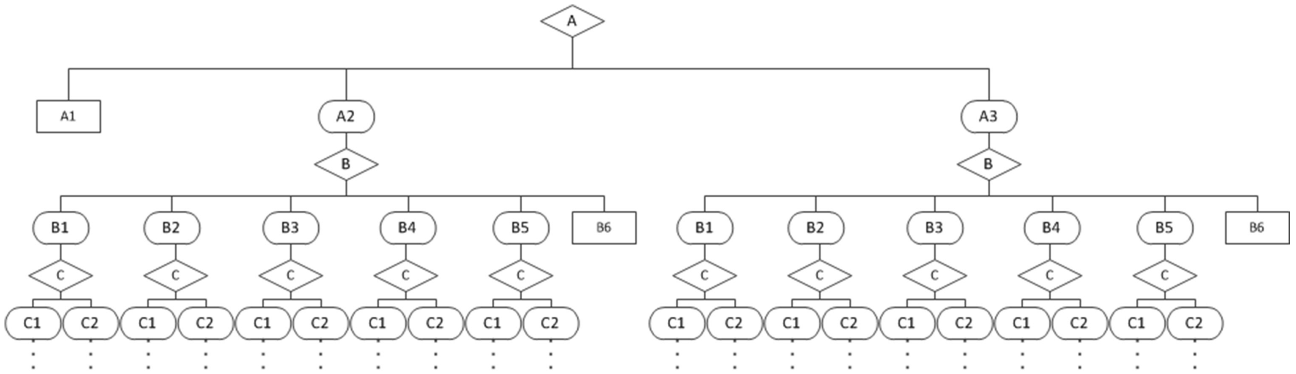
As an illustration of the STSS algorithm approach, a subsection of the pain management algorithm is shown in Figure 4. The diagram in Figure 4 illustrates three levels of decision nodes from the pain algorithm. The first level addresses patient-reported pain severity, which is categorized into three groups. The second level represents a patient's opioid use within the past 24 hours, which is categorized into six groups. The third level depicts a patient's creatinine clearance, which is categorized into two groups.

The STTS algorithm would follow every pathway to an end node while keeping a record of branches that had not yet been traversed (ie, the stack). After processing an end node (ie, a unique clinical decision pathway), the algorithm would then revisit the last node it had placed on the stack (ie, *pop* it off the stack) and then attempt to use this *popped* node's connections to find a new unique end node that had not yet been processed. Through this systematic traversal of the clinical algorithm, every possible pathway was identified and sample patient variables were set to specific values to ensure that each path would be traversed with every testing cycle. Since the *correct* recommendations that should result from every pathway were defined, and since every input data set could be paired with the anticipated output recommendations, comparison of the actual output recommendations with the expected recommendations led to the identification of errors in the algorithm logic. In the pain algorithm case, if the input variables were set with a pain score of 8 (node A3 in Figure 4), with an opioid use history of slow release opioids only (node B4), and with a normal kidney function (node C1), the algorithm should generate a recommendation to add immediate release opioids to the patient's pain control regimen.

Phase 3

In *Phase 3*, we evaluated clinical use of the system among patients and HCPs in the two clinics via observations, structured interviews, and questionnaires. Patients and HCPs meeting the same criteria employed in *Phase 1* were recruited to participate in a feasibility trial (details previously reported [34]). In the final six months of the trial, we conducted a user evaluation. Data were collected from patients and HCPs using observations of CDS system use, standardized questionnaires, and structured interviews. Descriptive statistics were used to analyze quantitative data on the Acceptability E-Scale and scoring with preset item threshold of 4 on the 1-5 response scale. Qualitative data were content analyzed using NVivo 9.0 software [39].

Figure 4. Sample of stack-traversal tree-spanning algorithm approach.



Results

Phase 1

13 patients participated in two rounds of testing **:** four were from DFCI and nine were from BMC. The sample was 62% (8/13) female, 82% (9/11; 2 missing) had less than a high school education, 54% (7/13) reported minority race, 46% (6/13) reported that they never or rarely used computers, and median age was 57 years. The usability testing scores for all patients in both rounds exceeded minimal acceptability scores. The mean scores for all items of the Acceptability E-Scale were >4 and composite acceptability scores averaged 27.5 for *round 1* and 26.9 for *round 2*, which exceeded the predefined threshold of 24. Based on patient comments, design features were tailored to accommodate computer use in older adults who were acutely ill, assist low literacy adults with no previous use of computers, ensure understanding of the time anchors, and enable accurate data collection regarding self-report of symptoms and medication use.

Five HCPs participated in two rounds of usability sessions: four were from DFCI and one was from BMC. The sample was 80% (4/5) male, 60% (3/5) white, and median age was 49 years. The average usability testing subscale scores for participants ranged from 3.3 to 3.9 in *round 1* and 3.6 to 3.9 in *round 2*. Based on HCP comments, design features were tailored to ensure that summary reports were easy to read in a busy clinical setting, confirm that no extra work was required to access the forms, ensure that decision support was timely so that it could be used during the clinical visit, and guarantee that all of the symptoms that were assessed using the algorithms were displayed and easily seen by the HCPs.

Phase 2 - Testing Results

Unit testing required an estimated 240 person-hours over nine months, including time between rounds for corrections. The

XSL•FO
RenderX

simplest algorithm (fatigue) required the least testing: four rounds over eight weeks. The most complex algorithm (pain) included the adjustment and conversion of opioid doses, recommendations of specific doses of medications for neuropathic and somatic pain, and addition of bowel regimens. The pain algorithm required five rounds of testing over six months. The results of unit testing identified both runtime and logical programming errors prior to clinical application. Examples of logical errors discovered during unit testing included morphine equivalent dosing irregularities and flow chart/algorithm wording that necessitated clarification for correct representation in programming logic. A small number of problems related to incomplete reasoning or inconsistent recommendations of the original algorithms were also identified, such as the potentially confusing simultaneous recommendations to increase a medication for depression but maintain the medication for anxiety. Clinical experts defined solutions in these cases. After each revision to an algorithm and programmed rule, each algorithm was tested again to ensure adequacy of the revision until no further errors were identified.

Integration testing was conducted in twelve rounds over eight weeks, using 10 test cases, and required an estimated 40 person-hours. Most identified errors were due to incorrect submission of data from the clinical site, such as an *unevaluable date* format. Integration testing also identified errors in display of data on the clinician report, and was used to define final requirements for the report. Each error was addressed and tested in the subsequent round until no further errors were identified. At the end of the testing, the rules were 100% accurate once all errors were corrected.

Using the STTS method described above, we generated all possible combinations of data parameters and variable values to enable validation of the five complex symptom management algorithms. Two illustrative sets of paired data input parameters, and their corresponding recommendation outputs, are shown in Table 2.

Phase 3

43 patients (100% of those invited) participated in the evaluation: 42 were from DFCI and one was from BMC. The sample was 58% (25/43) female, and 95% (40/43; 1 missing) white, had a median age of 60, with 70% (30/43) reporting some college education, and 72% (31/43) reporting using computers often or very often. Participants completed the symptom self-report in an average of seven minutes, with the most common technical problem being timing out from the waiting room wireless connection, while medication review took less than two minutes on average. Average acceptability item scores for SAMI-L ranged from 4.21 to 4.98 (on a 1-5 scale). The average total score for the acceptability scale was 28 (of 30), exceeding the predefined threshold of 24 for acceptability. Most patients (58%, 25/43) would prefer assessments at every clinic visit, versus a greater or lesser frequency (16%, 7/43) or no preference (26%, 11/43). The majority of participants (72%, 31/43) preferred completing assessments during clinic visits, versus at home (12%, 5/43) or no preference (16%, 7/43), because it gave them something to do while waiting and was a more reliable way to ensure completion of the report. Facilitators for use included: improved communication with providers, having time to reflect on symptoms before the visit, helping pinpoint problems, and ease of use. The main barrier to use was unclear or limited options on SAMI-L questionnaires. Patients suggested having open-ended questions to identify additional issues of concern.

13 of 14 (93%) HCP participants randomized to the intervention arm participated in the evaluation: 11 HCPs were observed in 42 instances of receiving a SAMI-L report, and 13 HCPs completed structured interviews and usability questionnaires. HCP participants included seven physicians and six nurse practitioners. The sample was 54% (7/13) male, with median age of 40, and had a median of 12 years of experience in oncology. In 79% (33/42) of observations, HCPs received the report on average 21 minutes before the visit and took <1 minute to review the report. Usability scores for the report ranged from an average of 3.2 for usefulness to 4.5 for organization (on a 1-5 scale). Two-thirds of HCPs (9/13) reported using the algorithm-derived recommendations for pain most often, and those for dyspnea the least. Management of dyspnea was perceived as complex, and algorithm suggestions were seen as being too generic. Another barrier identified was lack of integration of the report into the flow of care. Facilitators of use were the reports' colorful scales and line graphs used for tracking symptoms. Calculations for opioid dosing, identification of patient distress, and suggestions for managing fatigue and opioid-induced constipation were perceived as helpful.

XSL•FO
RenderX

Table 2. Pairing of input data parameters with resultant recommendations

Input Data Parameters	Resultant Recommendations
• Pain self-report=6 (moderate) • Intermittent pain • Pain is achy, sharp, or in one spot • No current opioid medications • Serum creatinine=0.9 • Sex=male • Age=67 • Weight=84 kg • Platelets=183,000/mL • No history of gastrointestinal bleed • Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events bowel score=0 (no constipation) • No current bowel medications	• Give morphine sulfate Immediate Release 7.5-15 mg by mouth every 4 hours as needed • OR oxycodone 5-10 mg by mouth every 4 hours as needed • OR hydromorphone 2-4 mg by mouth every 4 hours as needed • Give acetaminophen 1000 mg by mouth three times a day for somatic pain NOT to exceed 3000 mg per day • OR ibuprofen 400 mg by mouth three times a day for somatic pain with omeprazole or pantoprazole 20 mg by mouth daily for GI protection • Suggest giving senna 1-2 tablets twice a day, up to a maximum of 4 tablets twice a day, AND docusate sodium 1 tablet twice a day, for prevention or treatment of opioid-induced constipation
• Pain self-report=8 (severe) • Constant pain • Pain is burning or shooting • Current opioid dose prescribed: oxycodone immediate release 15 mg by mouth, every 4 hours as needed, and oxycodone extended release 60 mg, by mouth twice a day, with actual past 24-hour oxycodone use equal to the maximum dose of 90 mg immediate release and 120 mg extended release • Serum creatinine=2.3 • Sex=female • Age=53 • Weight=61 kg • Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events bowel score=3 (moderate, Grade 2 constipation) • Taking sennosides, United States Pharmacopeia 17 mg, 2 tablets, twice a day	• For symptom relief, give oxycodone at 30 mg by mouth. If pain >7 after 1 hour, suggest a palliative care consult. If pain is <6 after 1 hour, suggest you use one of the following combinations of sustained release and rescue dose opioids. Adjust to available formulations. • If oxycodone sustained release preferred: ○ Give oxycodone sustained release 120 mg by mouth twice a day, OR 80 mg by mouth three times a day ○ Give oxycodone immediate release 30 mg by mouth every 4 hours as needed • If transdermal fentanyl patch preferred: ○ Give transdermal patch 175 mcg/hr ○ Give oxycodone immediate release 30 mg by mouth every 4 hours as needed OR hydromorphone immediate release 15 mg by mouth every 4 hours as needed. • Suggest giving gabapentin 100 mg by mouth twice a day from days 1-7, then 200 mg by mouth twice a day from days 8-28, for neuropathic pain • If ineffective after 28 days: discontinue gabapentin and give pregabalin 50 mg by mouth twice a day from days 1-7, increasing to 75-100 mg by mouth twice a day from days 8-28. • If pregabalin ineffective after 28 days, call palliative care consult • Suggest titrating current 2.0 sennosides tablets by mouth twice a day, up to a maximum 4 tablets by mouth twice a day, to reach goal bowel function of either 1 bowel movement per day or 1 bowel movement every other day • AND give milk of magnesia 30mL once daily OR dulcolax 10 mg by mouth or by rectum once daily OR miralax 17 g once daily

Discussion

In this paper we described the development and testing of a CDS system, the SAMI-L, that used complex algorithms to address the simultaneous management of five distressing symptoms in lung cancer patients. In previous studies, CDS was used to identify the presence of a single symptom using an algorithm with less than a dozen decision nodes that generated general recommendations [12,22], whereas the algorithms that were developed and tested in this project focused on five symptoms that contained decision nodes that varied from 29 for fatigue to 1425 for pain. Thus, the algorithms developed for this study were complex due to the number of symptoms addressed, and the number of decision nodes was large compared to previous studies. The complexity of these algorithms required a novel and rigorous approach to testing. The CDS system was acceptable and useful for patients and HCPs in preclinical and clinical settings.

The successful deployment of SAMI-L advances the field by demonstrating that complex clinical algorithms can be invoked in rule-based CDS systems to generate detailed patient-specific recommendations for use in the management of multiple symptoms at the point-of-care using patient-entered data. Most previously reported rule-based CDS systems have contained fewer than a dozen decision nodes and required only a small number of data parameters to function [40-42].

While SAMI-L provides an example for increasing the logic complexity of rule-based CDS systems, we recognize that SAMI-L represents only one approach to CDS (ie, CDS driven

XSL•FO
RenderX

by explicit care algorithms) and that other approaches exist for CDS that manage even greater levels of complexity. Perhaps the most complex CDS tool described to date is the Watson technology developed by IBM [43]. In contrast to the defined rules of SAMI-L, Watson uses sophisticated natural language processing, and powerful information mining and retrieval capabilities to provide clinical guidance [44]. Watson-enabled CDS may reflect one future direction for CDS; however, we maintain that there is still a role for CDS systems that facilitate adherence to defined evidence-based best practices, as shown with SAMI-L. Rule-based CDS can be built with currently available technology in areas for which guidelines are available. As long as boundaries are clearly defined (eg, normal renal function in SAMI-L), rule-based CDS can be robust and promote guideline adherence.

In addition to the high-powered information mining and retrieval CDS approach enabled by Watson, another approach to enable complex CDS includes supervised learning models. While these approaches are able to support complex decisions, they require large sets of labeled data for algorithm training, often lack generalizability, are difficult to ensure replicability, and are not always able to provide the rationale for CDS recommendations.

Within the domain of CDS for symptom management, SAMI-L advances the field by supporting simultaneous management of multiple distressing symptoms in patients with lung cancer, in contrast to most previously reported systems that focus on a single symptom or problem [15,16]. The SAMI-L system also incorporates a measurement-based approach using patient-reported symptom severity, age, comorbidities, laboratory values, and adherence to medications to instantiate symptom management algorithms that generate guidance for a report delivered to clinicians in real-time. To our knowledge, this system is the first to provide CDS for management of multiple symptoms in oncology.

Another important facet of the SAMI-L system is that it produced immediate CDS for cancer symptom management based on complex logic utilizing patient data entered in real-time. The real-time collection of current symptom status from patients enabled SAMI-L to be responsive to the immediate needs of patients. The CDS tool was able to provide explicit advice for medication initiation or adjustment, as well as other interventions at the point-of-care. Enabling CDS to be responsive to current patient needs will become increasingly important as more data are collected in real-time through advances in patient-centric technologies.

From a technology standpoint, we validated the Web service approach for disassociating the collection of data and use of recommendations (in Massachusetts) from the decision engine (initially hosted on local servers in North Carolina and later moved to a cloud-based service). This project demonstrates that the client application can be separated from the decision engine over significant distances without compromising performance. The consistent function of SAMI-L demonstrates that Web service performance readily supports real-time, production-level-use CDS applications that deliver recommendations into workflow at the point-of-care. The Web service model would also accommodate potential reuse of the decision logic and scaling of the number of clients. As Dixon et al [45] note, provision of CDS by Web service opens the door to support for clinicians in settings with limited resources. Similar to the Dixon et al study, the SAMI-L decision engine could receive data from, and return decision support to, nonaffiliated health systems using secure protocols. Steurbaut et al [46] cite reduction of work overload as an additional advantage of a Web service approach.

This paper also demonstrates the magnitude of the testing required when implementing CDS using complex algorithms with over a hundred decision nodes and hundreds of possible values for the algorithm variables. The net result was more than a million possible unique data-parameter sets for traversing the most complex algorithm. The increased complexity of the logic supported by the SAMI-L CDS system necessitated new approaches to CDS testing. By using the STTS approach, we validated five complex CDS protocols for symptom management in cancer patients. In order to verify the accuracy of each algorithm, we automated the creation of hundreds of test data sets that enabled the assessment of boundary conditions, as well as the changing of multiple variables simultaneously. Thus, the STTS approach enabled boundary testing that would have otherwise been nearly impossible to achieve through a manual process, due to the protocol complexity. Moreover, this approach accommodated iterative testing of each protocol as it was refined by clinical experts, and allowed the testing process to be independent of the decision engine and the care protocol. In terms of generalizability, the testing framework used to validate SAMI-L can serve as a general model for testing CDS systems driven by complex algorithms in any clinical domain. In addition to the STTS approach, we manually constructed 10 sample cases derived from patients that reflected diverse symptomology, in order to test the entire system using all algorithms. We used this set of 10 test-cases to reassess system performance when modifications were made to the decision logic, since the change in the output reflects only the logic change, leaving all other recommendations constant.

One unanticipated issue was that hundreds of hours were needed to validate the algorithms before clinical implementation. In addition, a more iterative and user-centered design process between clinicians, research staff, and computer developers would have been ideal throughout the algorithm development cycle [47]. The expert panels produced algorithm flowcharts at the end of their work, and then programming of the decision rules began. When questions arose during programming, the expert panels were no longer meeting, and we had *ad hoc* access to only two clinical experts (palliative care and psychiatry), which created a slow and limited ability to address issues that arose during programming.

In terms of future directions for this work, SAMI-L should be tested in multiple clinics and used for symptom management for other types of cancer, especially in settings that have limited access to palliative care services [22]. In addition, the portability and shareability of the CDS logic via Web services should be demonstrated by allowing other CDS systems to use components of the SAMI-L knowledge base with new client applications.

XSL•FO
RenderX

Limitations

One limitation of the current study is that we used paper copies of reports rather than integrating the system into the electronic health record (EHR) system. This approach was necessary, given the feasibility nature of the study, the need to establish efficacy of the technology, and the high cost of integrating the system within the EHR. It was important not to disrupt usual workflow, so our research staff worked collaboratively with the clinical staff to make sure the reports were readily accessible to HCPs prior to the clinic visits. Future studies that test the efficacy of this approach should explore mechanisms to integrate the technology into the EHR, ensuring that this approach has the potential to be broadly applied if efficacious.

Conclusions

Complex algorithms can be invoked through rule-based CDS systems to promote evidence-based care in real-time at the point of patient contact using current, patient-supplied information to generate explicit, detailed, and patient-specific care guidance. This information collected in real-time from patients can be used to inform the symptom management process and serve to prioritize management interventions.

The increasing complexity of rule-based CDS systems requires new approaches to conduct thorough testing and validation of CDS systems, such as the STTS algorithm utilized in this project. Web services using a cloud-based decision engine can support clinical use of a CDS tool, in which the client application is independent and separate from the CDS engine.

Acknowledgments

The R01 CA125256 grant was prepared as part of a Mentored Career Development Award (1 K07 CA92696 - ME Cooley); Karen M Emmons, PhD and Bruce E Johnson, MD acted as mentors. This study was funded in part by a grant from the National Cancer Institute (5R01 CA125256), and the Duke University Department of Community and Family Medicine.

Authors' Contributions

DFL led the informatics aspect of this project, and participated in conception of the project, development and testing of the algorithms, implementation of the intervention, analyses and interpretation of data, writing, revising, and final approval of the manuscript. EBJ provided clinical informatics expertise, and participated in programming and testing the algorithms, implementation of the intervention, writing, revising, and approval of the manuscript. BH was the project director and participated in testing the algorithms, implementation of the intervention, data collection, analyses and interpretation of data, writing, revising, and final approval of the manuscript. TAS participated in testing of the algorithms, and implementation of the intervention, data collection, writing, revising, and final approval of the manuscript. JB participated in the informatics aspect of the project, and implementation of the intervention, writing, revising, and final approval of the manuscript. GDF provided clinical informatics expertise, and participated in programming the algorithms, writing, revising, and approval of the manuscript. DLB participated in conception of the project, and implementation of the intervention, writing, revising, and final approval of the manuscript. IMB provided clinical expertise to the project, and assisted in development of the algorithms, implementation of the intervention, writing, revising, and final approval of the manuscript. KF participated in conception of the project, and implementation of the intervention, writing, revising, and final approval of the manuscript. JW participated in the informatics part of the study, and implementation of the intervention, writing, revising, and final approval of the manuscript. JLA participated in conception of the project, and development and testing of algorithms, implementation of the intervention, analyses and interpretation of data, writing, revising, and final approval of the manuscript. MEC participated in the conception of the project, and the development and testing of the algorithms, implementation of the intervention, data collection, analyses and interpretation of data, writing, revising, and final approval of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Data requirements for the Symptom Assessment and Management Intervention (SAMI-L) System.

[PDF File (Adobe PDF File), 33KB - medinform_v4i4e36_app1.pdf]

Multimedia Appendix 2

Symptom management guidance and required data elements for the algorithms.

[PDF File (Adobe PDF File), 28KB - medinform_v4i4e36_app2.pdf]

References

XSL•FO
RenderX

1. 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]
2. Wagholikar KB, MacLaughlin KL, Kastner TM, Casey PM, Henry M, Greenes RA, et al. Formative evaluation of the accuracy of a clinical decision support system for cervical cancer screening. J Am Med Inform Assoc 2013;20(4):749-757 [FREE Full text] [doi: 10.1136/amiajnl-2013-001613] [Medline: 23564631]
3. Trafton JA, Martins SB, Michel MC, Wang D, Tu SW, Clark DJ, et al. Designing an automated clinical decision support system to match clinical practice guidelines for opioid therapy for chronic pain. Implement Sci 2010;5:26 [FREE Full text] [doi: 10.1186/1748-5908-5-26] [Medline: 20385018]
4. Sesen MB, Peake MD, Banares-Alcantara R, Tse D, Kadir T, Stanley R, et al. Lung Cancer Assistant: a hybrid clinical decision support application for lung cancer care. J R Soc Interface 2014 Sep 6;11(98):20140534 [FREE Full text] [doi: 10.1098/rsif.2014.0534] [Medline: 24990290]
5. Degner LF, Sloan JA. Symptom distress in newly diagnosed ambulatory cancer patients and as a predictor of survival in lung cancer. J Pain Symptom Manage 1995 Aug;10(6):423-431. [Medline: 7561224]
6. Kurtz ME, Kurtz JC, Stommel M, Given CW, Given BA. Symptomatology and loss of physical functioning among geriatric patients with lung cancer. J Pain Symptom Manage 2000 Apr;19(4):249-256. [Medline: 10799791]
7. Kurtz ME, Kurtz JC, Stommel M, Given CW, Given B. Predictors of depressive symptomatology of geriatric patients with lung cancer-a longitudinal analysis. Psychooncology 2002;11(1):12-22. [Medline: 11835589]
8. Tishelman C, Petersson L, Degner LF, Sprangers MA. Symptom prevalence, intensity, and distress in patients with inoperable lung cancer in relation to time of death. J Clin Oncol 2007 Dec 1;25(34):5381-5389 [FREE Full text] [doi: 10.1200/JCO.2006.08.7874] [Medline: 18048819]
9. Lemonnier I, Guillemin F, Arveux P, Clément-Duchêne C, Velten M, Woronoff-Lemsi M, et al. Quality of life after the initial treatments of non-small cell lung cancer: a persistent predictor for patients' survival. Health Qual Life Outcomes 2014;12:73 [FREE Full text] [doi: 10.1186/1477-7525-12-73] [Medline: 24884836]
10. Cooley ME, Short TH, Moriarty HJ. Symptom prevalence, distress, and change over time in adults receiving treatment for lung cancer. Psychooncology 2003;12(7):694-708. [doi: 10.1002/pon.694] [Medline: 14502594]
11. Hopwood P, Stephens RJ. Depression in patients with lung cancer: prevalence and risk factors derived from quality-of-life data. J Clin Oncol 2000 Feb;18(4):893-903. [Medline: 10673533]
12. Wang XS, Fairclough DL, Liao Z, Komaki R, Chang JY, Mobley GM, et al. Longitudinal study of the relationship between chemoradiation therapy for non-small-cell lung cancer and patient symptoms. J Clin Oncol 2006 Sep 20;24(27):4485-4491 [FREE Full text] [doi: 10.1200/JCO.2006.07.1126] [Medline: 16983118]
13. Boyes AW, Girgis A, D'Este CA, Zucca AC, Lecathelinais C, Carey ML. Prevalence and predictors of the short-term trajectory of anxiety and depression in the first year after a cancer diagnosis: a population-based longitudinal study. J Clin Oncol 2013 Jul 20;31(21):2724-2729 [FREE Full text] [doi: 10.1200/JCO.2012.44.7540] [Medline: 23775970]
14. Koczywas M, Cristea M, Thomas J, McCarty C, Borneman T, Del Ferraro C, et al. Interdisciplinary palliative care intervention in metastatic non-small-cell lung cancer. Clin Lung Cancer 2013 Nov;14(6):736-744 [FREE Full text] [doi: 10.1016/j.cllc.2013.06.008] [Medline: 23871439]
15. Cleeland CS, Portenoy RK, Rue M, Mendoza TR, Weller E, Payne R, et al. Does an oral analgesic protocol improve pain control for patients with cancer? An intergroup study coordinated by the Eastern Cooperative Oncology Group. Ann Oncol 2005 Jun;16(6):972-980 [FREE Full text] [doi: 10.1093/annonc/mdi191] [Medline: 15821119]
16. Passik SD, Kirsh KL, Theobald D, Donaghy K, Holtsclaw E, Edgerton S, et al. Use of a depression screening tool and a fluoxetine-based algorithm to improve the recognition and treatment of depression in cancer patients. A demonstration project. J Pain Symptom Manage 2002 Sep;24(3):318-327. [Medline: 12458113]
17. Greer JA, MacDonald JJ, Vaughn J, Viscosi E, Traeger L, McDonnell T, et al. Pilot study of a brief behavioral intervention for dyspnea in patients with advanced lung cancer. J Pain Symptom Manage 2015 Dec;50(6):854-860. [doi: 10.1016/j.jpainsymman.2015.06.010] [Medline: 26166181]
18. Dean GE, Abu SE, Yingrengreung S, Ziegler P, Chen H, Steinbrenner LM, et al. Sleeping with the enemy: sleep and quality of life in patients with lung cancer. Cancer Nurs 2015;38(1):60-70. [doi: 10.1097/NCC.0000000000000128] [Medline: 25486204]
19. Yoong J, Traeger LN, Gallagher ER, Pirl WF, Greer JA, Temel JS. A pilot study to investigate adherence to long-acting opioids among patients with advanced lung cancer. J Palliat Med 2013 Apr;16(4):391-396. [doi: 10.1089/jpm.2012.0400] [Medline: 23445248]
20. Partridge AH, Seah DS, King T, Leighl NB, Hauke R, Wollins DS, et al. Developing a service model that integrates palliative care throughout cancer care: the time is now. J Clin Oncol 2014 Oct 10;32(29):3330-3336. [doi: 10.1200/JCO.2013.54.8149] [Medline: 25199756]
21. Smith TJ, Temin S, Alesi ER, Abernethy AP, Balboni TA, Basch EM, et al. American Society of Clinical Oncology provisional clinical opinion: the integration of palliative care into standard oncology care. J Clin Oncol 2012 Mar 10;30(8):880-887 [FREE Full text] [doi: 10.1200/JCO.2011.38.5161] [Medline: 22312101]

XSL•FO
RenderX

22. Nekhlyudov L, Levit L, Hurria A, Ganz PA. Patient-centered, evidence-based, and cost-conscious cancer care across the continuum: translating the Institute of Medicine report into clinical practice. CA Cancer J Clin 2014;64(6):408-421 [FREE Full text] [doi: 10.3322/caac.21249] [Medline: 25203697]
23. Cooley ME, Lobach DF, Johns E, Halpenny B, Saunders T, Del Fiol G, et al. Creating computable algorithms for symptom management in an outpatient thoracic oncology setting. J Pain Symptom Manage 2013 Dec;46(6):911-924.e1 [FREE Full text] [doi: 10.1016/j.jpainsymman.2013.01.016] [Medline: 23680580]
24. McCorkle R, Young K. Development of a symptom distress scale. Cancer Nurs 1978 Oct;1(5):373-378. [Medline: 250445]
25. Basch E, Reeve BB, Mitchell SA, Clauser SB, Minasian LM, Dueck AC, et al. Development of the National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). J Natl Cancer Inst 2014 Sep;106(9) [FREE Full text] [doi: 10.1093/jnci/dju244] [Medline: 25265940]
26. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [Medline: 11556941]
27. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983 Jun;67(6):361-370. [Medline: 6880820]
28. Kawamoto K, Lobach DF. Design, implementation, use, and preliminary evaluation of SEBASTIAN, a standards-based Web service for clinical decision support. In: AMIA Annu Symp Proc. 2005 Presented at: AMIA Annual Symposium; October 22-26, 2005; Washington, DC p. 380-384 URL: http://europepmc.org/abstract/MED/16779066
29. Kawamoto K, Del FG, Lobach DF, Jenders RA. Standards for scalable clinical decision support: need, current and emerging standards, gaps, and proposal for progress. Open Med Inform J 2010;4:235-244 [FREE Full text] [doi: 10.2174/1874431101004010235] [Medline: 21603283]
30. Borbolla D, Otero C, Lobach DF, Kawamoto K, Gomez Saldaño AM, Staccia G, et al. Implementation of a clinical decision support system using a service model: results of a feasibility study. Stud Health Technol Inform 2010;160(Pt 2):816-820. [Medline: 20841799]
31. Griebel L, Prokosch H, Köpcke F, Toddenroth D, Christoph J, Leb I, et al. A scoping review of cloud computing in healthcare. BMC Med Inform Decis Mak 2015;15:17 [FREE Full text] [doi: 10.1186/s12911-015-0145-7] [Medline: 25888747]
32. Given C, Given B, Rahbar M, Jeon S, McCorkle R, Cimprich B, et al. Effect of a cognitive behavioral intervention on reducing symptom severity during chemotherapy. J Clin Oncol 2004 Feb 1;22(3):507-516 [FREE Full text] [doi: 10.1200/JCO.2004.01.241] [Medline: 14752074]
33. Berger AM, Yennu S, Million R. Update on interventions focused on symptom clusters: what has been tried and what have we learned? Curr Opin Support Palliat Care 2013 Mar;7(1):60-66. [doi: 10.1097/SPC.0b013e32835c7d88] [Medline: 23364298]
34. Cooley ME, Blonquist TM, Catalano PJ, Lobach DF, Halpenny B, McCorkle R, et al. Feasibility of using algorithm-based clinical decision support for symptom assessment and management in lung cancer. J Pain Symptom Manage 2015 Jan;49(1):13-26 [FREE Full text] [doi: 10.1016/j.jpainsymman.2014.05.003] [Medline: 24880002]
35. Tariman JD, Berry DL, Halpenny B, Wolpin S, Schepp K. Validation and testing of the Acceptability E-scale for web-based patient-reported outcomes in cancer care. Appl Nurs Res 2011 Feb;24(1):53-58 [FREE Full text] [doi: 10.1016/j.apnr.2009.04.003] [Medline: 20974066]
36. Chin J, Diehl V, Norman K. Development of an instrument measuring user satisfaction of the human-computer interface. In: Human factors in computing systems: Consumer Health Informatics '88 conference proceedings. New York: Association for Computing Machinery; 1988 Presented at: SIGCHI Conference on Human Factors in Computing Systems; May 15-19, 1988; New York, NY URL: http://dl.acm.org/citation.cfm?doid=57167.57203 [doi: 10.1145/57167.57203]
37. Doll WJ, Torkzadeh G. The measurement of end-user computing satisfaction. MIS Quarterly 1988 Jun;12(2):259-258. [doi: 10.2307/248851]
38. Altova. MapForce graphical mapping software. Vienna: Altova; 2011. URL: https://www.altova.com/mapforce.html [accessed 2016-10-18] [WebCite Cache ID 6lMA0f4Od]
39. QSR International. QSR International Pty Ltd. 2010. NVivo qualitative data analysis Software, 9th ed URL: http://www.qsrinternational.com/ [accessed 2016-10-11] [WebCite Cache ID 6lBDr1WQ8]
40. Kurian BT, Trivedi MH, Grannemann BD, Claassen CA, Daly EJ, Sunderajan P. A computerized decision support system for depression in primary care. Prim Care Companion J Clin Psychiatry 2009;11(4):140-146 [FREE Full text] [doi: 10.4088/PCC.08m00687] [Medline: 19750065]
41. Rigopoulou AV, Anthracopoulos MB, Katsardis CV, Lymberopoulos DK. RespDoc: a new clinical decision support system for childhood asthma management based on Fraction of Exhaled Nitric Oxide (FeNO) measurements. Conf Proc IEEE Eng Med Biol Soc 2013;2013:1306-1309. [doi: 10.1109/EMBC.2013.6609748] [Medline: 24109935]
42. Peleg M, Fox J, Patkar V, Glasspool D, Chronakis I, South M, et al. A computer-interpretable version of the AACE, AME, ETA medical guidelines for clinical practice for the diagnosis and management of thyroid nodules. Endocr Pract 2014 Apr;20(4):352-359. [doi: 10.4158/EP13271.OR] [Medline: 24246343]
43. Kohn MS, Sun J, Knoop S, Shabo A, Carmeli B, Sow D, et al. IBM's health analytics and clinical decision support. Yearb Med Inform 2014;9:154-162 [FREE Full text] [doi: 10.15265/IY-2014-0002] [Medline: 25123736]

XSL•FO
RenderX

44. Upbin B. Forbes. 2013 Feb 08. IBM's Watson gets its first piece of business in healthcare URL: http://www.forbes.com/sites/bruceupbin/2013/02/08/ibms-watson-gets-its-first-piece-of-business-in-healthcare/#eaa2c4f44b16 [accessed 2016-10-17] [WebCite Cache ID 6lKwHBJq8]
45. Dixon BE, Simonaitis L, Perkins SM, Wright A, Middleton B. Measuring agreement between decision support reminders: the cloud vs. the local expert. BMC Med Inform Decis Mak 2014;14:31 [FREE Full text] [doi: 10.1186/1472-6947-14-31] [Medline: 24720863]
46. Steurbaut K, Van Hoecke S, Colpaert K, Lamont K, Taveirne K, Depuydt P, et al. Use of web services for computerized medical decision support, including infection control and antibiotic management, in the intensive care unit. J Telemed Telecare 2010;16(1):25-29. [doi: 10.1258/jtt.2009.001008] [Medline: 20086264]
47. Horsky J, Schiff GD, Johnston D, Mercincavage L, Bell D, Middleton B. Interface design principles for usable decision support: a targeted review of best practices for clinical prescribing interventions. J Biomed Inform 2012 Dec;45(6):1202-1216 [FREE Full text] [doi: 10.1016/j.jbi.2012.09.002] [Medline: 22995208]

Abbreviations

BMC: Boston Medical Center
CDS: clinical decision support
DFCI: Dana-Farber Cancer Institute
EHR: electronic health record
HCC: Health Communication Core
HCP: health care provider
SAMI-L: Symptom Assessment and Management Intervention for Lung cancer
SEBASTIAN: System for Evidence-Based Advice Through Simultaneous Transaction with an Intelligent Agent Across a Network
STTS: stack-traversal tree-spanning
XML: eXtensible Markup Language

Edited by G Eysenbach; submitted 08.03.16; peer-reviewed by A Wright, B Parker, L Griebel; comments to author 06.04.16; revised version received 16.08.16; accepted 03.09.16; published 08.11.16.

Please cite as:
Lobach DF, Johns EB, Halpenny B, Saunders TA, Brzozowski J, Del Fiol G, Berry DL, Braun IM, Finn K, Wolfe J, Abrahm JL, Cooley ME
Increasing Complexity in Rule-Based Clinical Decision Support: The Symptom Assessment and Management Intervention
JMIR Med Inform 2016;4(4):e36
URL: http://medinform.jmir.org/2016/4/e36/
doi:10.2196/medinform.5728
PMID:27826132

©David F Lobach, Ellis B Johns, Barbara Halpenny, Toni-Ann Saunders, Jane Brzozowski, Guilherme Del Fiol, Donna L Berry, Ilana M Braun, Kathleen Finn, Joanne Wolfe, Janet L Abrahm, Mary E Cooley. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 08.11.2016. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the Effect of Web-Based Iranian Diabetic Personal Health Record App on Self-Care Status and Clinical Indicators: Randomized Controlled Trial

Amirabbas Azizi[1*], PhD; Robab Aboutorabi[2*], Dr med; Zahra Mazloum-Khorasani[2*], Dr med; Monavar Afzal-Aghaea[3*], MD, PhD; Hamed Tabesh[4*], PhD; Mahmood Tara[4*], MD, PhD

[1]School of Paramedicine, Department of Health Information Technology, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Islamic Republic of Iran

[2]School of Medicine, Endocrine Research Center, Metabolic Syndrome Research Center, Mashhad University of Medical Sciences, Mashhad, Islamic Republic of Iran

[3]School of Health, Management & Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Islamic Republic of Iran

[4]School of Medicine, Department of Medical Informatics, Mashhad University of Medical Sciences, Mashhad, Islamic Republic of Iran

[*]all authors contributed equally

Corresponding Author:
Mahmood Tara, MD, PhD
School of Medicine
Department of Medical Informatics
Mashhad University of Medical Sciences
Pardis Daneshgah
Park Square
Mashhad, 917-7948-564
Islamic Republic of Iran
Phone: 98 5138002429
Fax: 98 5138002445
Email: taram@mums.ac.ir

Abstract

Background: There are 4 main types of chronic or noncommunicable diseases. Of these, diabetes is one of the major therapeutic concerns globally. Moreover, Iran is among the countries with the highest incidence of diabetic patients. Furthermore, library-based studies by researchers have shown that thus far no study has been carried out to evaluate the relationship between Web-based diabetic personal health records (DPHR) and self-care indicators in Iran.

Objective: The objective of this study is to examine the effect of Web-based DPHR on self-care status of diabetic patients in an intervention group as compared with a control group.

Methods: The effect of DPHR on self-care was assessed by using a randomized controlled trial (RCT) protocol for a 2-arm parallel group with a 1:1 allocation ratio. During a 4-month trial period, the control group benefited from the routine care; the intervention group additionally had access to the Web-based DPHR app besides routine care. During the trial, 2 time points at baseline and postintervention were used to evaluate the impact of the DPHR app. A sample size of 72 people was randomly and equally assigned to both the control and intervention groups. The primary outcome measure was the self-care status of the participants.

Results: Test results showed that the self-care status in the intervention group in comparison with the control group had a significant difference. In addition, the dimensions of self-care, including normal values, changes trend, the last measured value, and the last time measured values had a significant difference while other dimensions had no significant difference. Furthermore, we found no correlation between Web-based DPHR system and covariates, including scores of weight, glycated hemoglobin (HbA1c), serum creatinine, high-density lipoprotein (HDL), low-density lipoprotein (LDL), total cholesterol, and planned visit adherence, as well as the change trend of mean for blood glucose and blood pressure.

Conclusions: We found that as a result of the Web-based DPHR app, the self-care scores in the intervention group were significantly higher than those of the control group. In total, we found no correlation between the Web-based DPHR app and

XSL•FO
RenderX

covariates, including planned visit adherence, HbA1c, serum creatinine, HDL, LDL, total cholesterol, weight, and the change trend of mean for blood glucose and blood pressure.

ClinicalTrial: Iranian Registry of Clinical Trials (IRCT): 2013082914522N1; http://www.irct.ir/searchresult.php?id=14522&number=1 (Archived by WebCite at http://www.webcitation.org/6cC4PCcau)

(JMIR Med Inform 2016;4(4):e32) doi:10.2196/medinform.6433

KEYWORDS

systematic review; diabetes mellitus, type 2; health records, personal; electronic health records; self care; Iran

Introduction

There are 4 main types of chronic or noncommunicable diseases (NCDs) [1]. Of these, diabetes is one of the major therapeutic concerns globally [2-5]. According to the World Health Organization (WHO) [6], diabetes is a chronic disease that occurs when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin. Type 2 diabetes (formerly called noninsulin-dependent or adult-onset diabetes) is caused by ineffective use of insulin in the body.

In terms of improving the management of diabetes, efforts made to enhance the self-care status of diabetic patients are of utmost importance [7-9]. The management of chronic diseases such as diabetes mellitus (DM), in comparison with other chronic conditions, is heavily dependent on the individuals and regular assessment by the health care providers [10]. According to a report, approximately 90% of diabetics suffer from type 2 diabetes [11]. One of the most important concerns in the public health system is the medical care average cost of type 2 DM, which is almost 3 times more than others [12]. Therefore, improving self-care skills among individuals with chronic diseases will resolve many challenges to health systems.

Self-care behaviors refer to decisions a person can make and activities he or she can do to deal with a health issue or improve his or her health status. Self-care behaviors that people need to learn or improve in order to deal with type 2 diabetes effectively are self-monitoring of blood sugar, healthy diet, regular exercise, and adherence to medical treatment [13]. There are widely different models of self-care behaviors with the common feature in which the patient acts as the heart of health management. From the perspective of health promotion, health is taken into account as a source of daily life and self-care status is considered as empowerment. Thus, through the acquisition of self-care skills, people are able to actively get involved in decisions affecting their health [14].

It is recognized that the incidence of diabetes has been steadily rising for the past few decades around the world, especially with the highest rate growing more rapidly in middle- and low-income countries. According to the public call by WHO to cope with diabetes in all countries of the world together with its recognition as an alarm, especially in the developing countries, the management of such a disease by Iran's Ministry of Health is taken into account as one of the research priorities [15,16]. Moreover, Iran is among the countries with the highest incidence of diabetic patients [17].

The reports presented by the International Diabetes Federation (IDF) classification cover the 20 countries and territories of the IDF MENA (Middle East and North Africa) region including Iran. There are over 37 million diabetics in these regions among the 387 million subjects suffering from this disease throughout the world and it is expected to reach 68 million by 2035. The number of such patients was 4.5 million in Iran in 2014, and in the same year, the incidence rate of the disease was 8.6% and 9% [18] in Iran and the world, respectively.

The therapeutic care of diabetic patients is still suboptimal despite international efforts often due to the lack of patient interactions with health care providers that are toward Web-based interventions [19]. Web-based personal health records (PHRs) are e-tools that allow patients to access health information via the Internet and take a more active role in their own health [20-22]. Patient-centric nature of PHRs make them ideal for patients to switch paternalistic model of medical care to a patient-centered model in which the patient is motivated to be an active and informed member of the health care team [23]. A review of the related literature on PHR revealed that such research studies have been different in terms of the following aspects:

- The first difference was associated with PHR format (paper or electronic). The review of literature indicated that the majority of research studies across the world have been conducted on electronic PHR. One significant reason could be that in such cases, requirements and prerequisites of electronic PHR studies are available, for example, an Electronic Health Record (EHR) or Electronic Medical Record (EMR) system is installed and PHR information is linked to that system.
- The second difference was associated with the subjects covered by PHR research studies. In other words, PHR has been conducted in multiple health issues, predominantly related to chronic diseases, including diabetes, cancer, and preventive care.
- The third difference concerned the sample size of PHR research studies. Some studies have been carried out on a small sample size [24,25] and others on a very large one [26]. In this respect, researchers found that paper-based PHR studies encompassed a small sample size and electronic PHR studies had been conducted by employing a large sample size.
- Last but not least, the fourth difference was linked with the study design of PHR research studies. The majority of studies in the field of PHR have been conducted in a retrospective manner. Additionally, there are numerous studies, merely library-based, discussing the definitions proposed for PHR. However, there are several studies

XSL•FO
RenderX

evaluating PHR through quasi-experimental and randomized controlled trial (RCT).

There are numerous studies across the world, investigating the relationship between paper-based or electronic PHR and indicators such as self-care, self-efficacy, and quality of life as primary outcome measures. Moreover, in these studies, clinical indicators, including lipid profile, blood glucose, blood pressure, weight, and glycated hemoglobin (HbA1c) have been evaluated as secondary outcome measures [11,27-32]. Although there are many studies on the impact of PHR interventions on self-care index and clinical outcomes related to diabetic patients in the world, yet they are limited in developing countries. In addition, a similar study in Iran was conducted on the relationship between the use of paper-based diabetes follow-up card and self-management among diabetic patients [33]. Furthermore, library-based studies by researchers have shown that thus far no study has been carried out to appraise the relationship between Web-based DPHR and self-care indicators in Iran. According to previous studies, further research is needed to investigate the impact of electronic media on patient self-care behaviors [34,35]. The final DPHR model was systematically developed in our former study [36] and the purpose of this study is to evaluate the effect of the Web-based DPHR app on self-care status and clinical outcome measures. In this study, we hypothesized that the participants assigned to receive the Web-based DPHR app will manage better self-care compared with those who received usual care.

Methods

Study Overview

In the first phase, the initial version of the DPHR model was designed through systematic review and then validated and confirmed by the contribution of local endocrinologists. The details associated with the gray literature and databases, the quality appraisal of evidences, and the validation technique employed for DPHR through the Delphi method were mentioned in a review conducted by the authors of this study [36]. In addition, the details of the research method related to the second phase of this study are explained as follows:

Diabetic Personal Health Records App Development

Web-based DPHR app was coded through PHP programming language. Its server operating system was Linux and its database was MySQL. The app was developed by 2 professionals in this domain. In order to complete the DPHR development, 20 sessions were held for almost 200 hours. Two-type designed DPHR interface supported both the patients and the senior investigator (as a system administrator). SMS text messaging (short message service, SMS) and phone call were considered as reminders to check the DPHR system.

Web-based DPHR is a system by which type 2 diabetic patients can manage their health information associated with diabetes. The information in the app is obtained based on the systematic review of the valid references, including articles, reports, standards, and guidelines of international institutes. In sum, monitoring data, history of progress, appointment schedule, acquaintance with the disease; entering the health history, blood sugar levels, lab tests, blood pressure, weight, height and body mass index (BMI); and knowing the past and future time of medical advices and visits to improve self-awareness and self-care should be implemented easily through this app.

Usability Evaluation

Prior to the implementation of the app in a real context, the app interface was refined and optimized throughout the trial using heuristic usability evaluation techniques [37] by medical informaticians, endocrinologists, as well as using think-aloud technique by the participants. Moreover, a 3-part questionnaire was employed to elicit the views of the patients about the app. The components of the above-mentioned questionnaire were the general characteristics of the patients (including 10 data items), the user’s tasks (including 10 tasks), and the evaluation questions of the app (including 8 questions).

Functions of Diabetic Personal Health Records App

The functions of Web-based DPHR app are as follows:

- Identifying and maintaining a patient’s record: Through this function, users would be able to record and view the personal information, the urgent contact information, the diabetes information, comorbidities, the risk factors, and the allergy and vaccination information.
- Managing body mass index: Through this function, users would be able to record their weight and height, and subsequently body mass index is automatically calculated by the system.
- Managing lab tests: This function enables users to record their lab tests.
- Managing patient history: This function enables users to view all the information recorded and edit them if necessary.
- Managing patient visits: By employing this function, users would be able to view previous and future visits.
- Managing the physician’s advices: Through this function, users would be able to view the physician’s advices.
- Health dashboard: This section is one of the most important parts of the system by which the users could view the latest information in the form of graphs and view their status through the existing colors. For example, green represents normal status.

The research group and the app provider controlled development, updates, and maintenance of the DPHR system. The patients had additional interventions such as more visits and experimental tests, apart from prescribed therapeutic procedures of a physician to assess variations in the status of the health information.

Study Design

The effects of DPHR on self-care status were assessed by using a RCT protocol for a 2-arm parallel group with a 1:1 allocation ratio. During a 4-month period, the control group benefited from the usual care; the intervention group additionally had access to the Web-based DPHR app besides routine care. Also, 2 time points at baseline and postintervention were used to evaluate the impact of the DPHR app.

The members who participated in the trial and gave informed consent based on some parameters including sex (male, female),

XSL•FO
RenderX

employment status (employed and unemployed), and age ranges (≤30, 30-50, and ≥50 years), were randomly allocated in the 2 groups regarding covariate-adaptive randomization through SPSS version 21.0 (IBM Corp), by a person with no direct role in the research. A senior investigator and data analyst were blinded during the trial, unlike participants and practitioners who could not be blinded to DPHR since it was an obvious artifact. In addition, the individuals in both groups were not allowed to exchange DPHR information to avoid contamination of the trial.

Participants

The statistical population of this 4-month trial in 2015 included patients suffering from type 2 DM in one of the endocrinology practice offices in Mashhad city, where there are over 120,000 patients with diabetes [38] considering inclusion and exclusion criteria. In the given office, there was a medical record for each patient in which all patient referrals were documented in its related record. In this study, we used the data from the records, including the number of patients based on their disease types, and extracted the demographic profile to do the introductory studies.

Only participants with signed informed consent were included in the study and were randomly divided into 2 groups: intervention and control. It should be noted that they manually received a package including a copy of the consent form, welcome letter, take-home manual, and stepwise instructions of the app usage. Their communication modes, in order to pose questions and concerns with the trial assistant, were either phone or SMS text messaging. The trained assistant required information concerning the intervention process, the Web-based DPHR app and the possible questions and answers.

The study's inclusion criteria included: the age range of 20-70 years, resident of Mashhad, at least one-year history of having type 2 diabetes, knowledge of computers and access to the Internet, high school diploma or above, as well as completing an informed consent. The exclusion criteria included lack of cooperation or inability to perform the study for any reason such as sickness, pregnancy, immigration, and so on.

Sample Size

There was no previous research about self-care among the Iranian population to estimate the sample size, except a survey showing an association between self-care activities and the quality of life [39]. Thus, in this work the sample size of 60, corresponding to the formula, was estimated based on the same one in Iran [40]. Finally, 72 patients in the 2 groups, of whom 36 cases were from the intervention group receiving the DPHR app and 36 from the control group with the routine cares, were enrolled according to the confidence interval of 95%, the power of 80%, and the dropout rate of 20%.

Outcome Measures of Study

In this study there was one primary outcome and several secondary outcomes. To evaluate the self-care status as the primary outcome measure, a researcher-made questionnaire composed of 7 sections with independent items was developed. This questionnaire was adapted from the existing valid literature in the field of self-care [41-44]. Self-care is one of the measures related to knowledge used by patients and in fact, the reason for selection of this criterion [45]. The dimensions of the self-care questionnaire are as follows: general information (25 questions), information of normal values (10 questions), information of change trend (10 questions), information of physician advices (2 questions), visit information (3 questions), information of the latest measurement values (9 questions), information of date and time of the latest measurement values (9 questions), and information of training tips (13 questions). In addition, some secondary outcomes were observed in this study. Any potential relation compared with cause and effect was investigated between the use of the app and the diabetes follow-up clinical indicators as secondary outcome measures. Such indicators were: fasting blood sugar (FBS), 2-hour postprandial blood sugar, weight, blood pressure, lipid profile (total cholesterol, triglyceride, HDL, and LDL), HbA1c, and serum creatinine. Furthermore, another secondary outcome measure was the adherence of patient to planned visit. The outcome measures of this study is presented in Table 1.

XSL•FO
RenderX

Table 1. Outcome measures of study.

Outcome measures	Measurement time points		
	Baseline	Weekly	Postintervention
Primary			
Self-care status	X		X
Secondary			
Blood sugar: FBS 2-hour postprandial	X	X	X
Weight	X		X
Blood pressure: Systolic Diastolic	X	X	X
Lipid profile: Total cholesterol Triglyceride HDL LDL	X		X
HbA1c	X		X
Serum creatinine	X		X
Adherence to planned visit	X		X

Data Collection

We began the study by acquiring the following baseline information: gender, age, marital status, occupation, education level, family history of DM, types of drug used, history of high blood pressure, access to home monitoring tools (glucometer, sphygmomanometer, scale), computer literacy, access to the Internet, working time with a computer, length of disease, FBS, 2-hour postprandial blood sugar, blood pressure, weight, lipid profile (total cholesterol, triglyceride, HDL, and LDL), serum creatinine, and HbA1c along with the information required to check the inclusion eligibility.

The trial assistant informed the patients elected by the inclusion criteria about the details of the study and then obtained informed written consent, as well as the tool required including the questionnaire that was completely anonymous to respect the ethical considerations. Each form had a unique code to manage further references. They were free to contact for sharing the questions and concerns by available communication means during the entire project. The assistant provided a username and password to access the DPHR app, and data were recorded securely only by authorized members of the trial team.

The face and content validity of the self-care questionnaire was assessed by the experts' opinions (including 2 endocrinologists, 1 medical informatician, and 1 methodologist) and valid literature. Also, the questionnaires were completed through structured interviews by a trial assistant blinded in the study, and the participants completed the self-care questionnaire on paper in person at the diabetes clinic during the follow-up phase. The final score of the self-care index was obtained from the total correct responses for each item in the 7-part questionnaire. Each correct response was assigned a score of 1 and the wrong answers, a 0. The scores of each dimension of the questionnaire were obtained from the sum of correct answers to the items of that dimension.

Data related to HbA1c, lipid profile, and weight in pre- and post-intervention were gathered in both control and interventions groups, but weekly blood sugar and blood pressure measurements were done only in the intervention group where the patients were responsible for doing them on their own, including weight measures. Other tests such as HbA1c, lipid profile and serum creatinine were performed in a laboratory based on objective- and laboratory-based measures. Moreover, the reminder means to view the app or complete their self-care actions were through weekly SMS text messaging or phone call. Figure 1 depicts the flow diagram of the study procedures.

Data Analysis

The statistical significance of the 2-tailed analyses in this study was performed with a significance interval of 95% and alpha level was set at *P*<.05. The data were analyzed through descriptive analysis (frequency, percent, and mean) and inferential analysis (normality test, paired and unpaired t-test, chi-square test, Fisher's exact test, Mann-Whitney U test, and 1-way ANOVA), using SPSS version 21.0. The descriptive statistics analyzed distribution of variables, website usage statistics, and app visits. In case of analyzing the variables under review, initially the value of outcome measure (eg, self-care) was analyzed independently in both groups according to data of baseline and postintervention phases using the paired t-test analysis and then the scores of both groups were analyzed using the independent t-test analysis.

XSL•FO
RenderX

Ethical Considerations

This study, registered on Iranian Registry of Clinical Trials (IRCT), received approval from the Research Review Committee and the Regional Ethics Committee (approval # 921835). Moreover, the details of research protocol have been published [46].

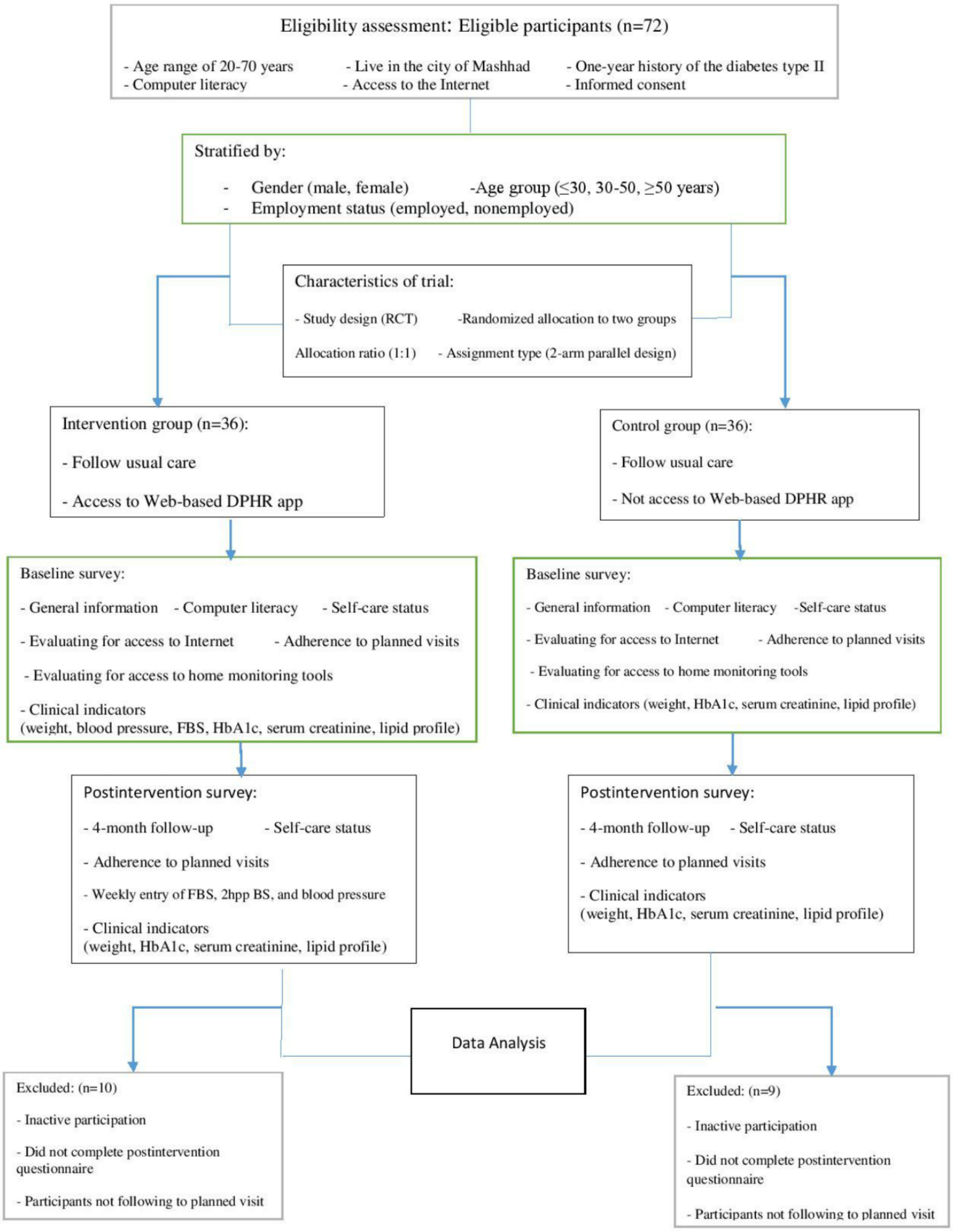
Results

App Usage

Statistics on DPHR app usage were obtained through webserver log analysis. The final analysis comprised 27 out of 36 participants in the control group and 26 out of 36 in the intervention group. Hence, the rates of patient response to the self-care questionnaire were 75% and 72% in the control and intervention groups, respectively.

According to Table 2, the maximum frequency of measurement record was 63 times, and the minimum value was 10 times. Blood glucose followed by blood pressure, weight, and lab tests were the most-recorded parameters during the study. In total, the highest record was related to blood sugar levels with a frequency of 450 times (mean 17.3), and the lowest one was associated with lab test entries with a frequency of 53 times (mean 2). Additionally, 38.5% (10/26) had recorded the measurement less than 20 times, 38.5% (10/26) 20-30 times, and 23% (6/26) over 30 times. The details related to the variables under review recorded by the patients in Web-based DPHR app are presented in Table 2.

Figure 1. Flow diagram of study procedures. DPHR: diabetic personal health record; FBS: fasting blood sugar; Hb^{A1c}: glycated hemoglobin; 2hpp BS: 2-hour post-prandial blood sugar.


Eligibility assessment: Eligible participants (n=72)
- Age range of 20-70 years
- Computer literacy
- Live in the city of Mashhad
- Access to the Internet
- One-year history of the diabetes type II
- Informed consent
Stratified by:
- Gender (male, female)
-Age group (≤30, 30-50, ≥50 years)
- Employment status (employed, nonemployed)
Characteristics of trial:
- Study design (RCT)
-Randomized allocation to two groups
Allocation ratio (1:1)
- Assignment type (2-arm parallel design)
Intervention group (n=36):
- Follow usual care
- Access to Web-based DPHR app
Control group (n=36):
- Follow usual care
- Not access to Web-based DPHR app
Baseline survey:
- General information
- Computer literacy
- Self-care status
- Evaluating for access to Internet
- Adherence to planned visits
- Evaluating for access to home monitoring tools
- Clinical indicators
(weight, blood pressure, FBS, HbA1c, serum creatinine, lipid profile)
Baseline survey:
- General information
- Computer literacy
-Self-care status
- Evaluating for access to Internet
- Adherence to planned visits
- Evaluating for access to home monitoring tools
- Clinical indicators (weight, HbA1c, serum creatinine, lipid profile)
Postintervention survey:
- 4-month follow-up
- Self-care status
- Adherence to planned visits
- Weekly entry of FBS, 2hpp BS, and blood pressure
- Clinical indicators
(weight, HbA1c, serum creatinine, lipid profile)
Postintervention survey:
- 4-month follow-up
- Self-care status
- Adherence to planned visits
- Clinical indicators
(weight, HbA1c, serum creatinine, lipid profile)
Data Analysis
Excluded: (n=10)
- Inactive participation
- Did not complete postintervention questionnaire
- Participants not following to planned visit
Excluded: (n=9)
- Inactive participation
- Did not complete postintervention questionnaire
- Participants not following to planned visit

XSL•FO
RenderX

Table 2. The frequency and average of trial variable entries in diabetic personal health records (DPHR) app (n=26).

Patient ID	Weight and height	Blood pressure	Lab tests	Blood sugar	Total
1	2	3	2	8	15
2	2	2	2	11	17
3	4	8	4	10	26
4	2	4	2	12	20
5	3	18	2	40	63
6	2	4	2	16	24
7	3	1	2	9	15
8	1	3	2	10	16
9	2	8	2	41	53
10	2	4	2	8	16
11	4	7	2	15	28
12	4	10	2	40	56
13	2	4	2	6	14
14	2	3	2	36	43
15	2	4	2	14	22
16	3	4	2	10	19
17	2	10	2	24	38
18	1	3	2	4	10
19	3	4	2	37	46
20	2	3	2	8	15
21	2	6	1	15	24
22	2	3	2	15	22
23	2	2	2	20	26
24	2	3	2	13	20
25	1	3	2	16	22
26	2	3	2	12	19
Total	59	127	53	450	689
Mean	2.2	4.8	2	17.3	26.5

Table 3. The demographic profile and opinions of diabetic patients participating in usability evaluation of diabetic personal health records (DPHR) app (n=6).

No.	Gender	Age (year)	Employment status	Education level	Time (minute)
1	Female	50	Unemployed	BSc	40
2	Female	58	Employed	BSc	20
3	Male	36	Employed	BSc	30
4	Male	38	Employed	BSc	25
5	Female	58	Unemployed	Diploma	20
6	Male	61	Unemployed	Diploma	45

XSL•FO
RenderX

Table 4. The opinions of experts and diabetic patients participating in usability evaluation of diabetic personal health records (DPHR) app.

Target groups	Opinions
Medical informatics students	Systolic blood pressure should be written in Persian.
	It would be better to use the full screen, especially in browsing the page in order not to scroll so much.
	The fonts of data entry forms were different from those of report forms.
	In the graphs of blood sugar history, the values could be shown in green or red for the normal and abnormal ranges in order to help patients know their status.
	Charts in the main page were incomprehensible.
	The first and the last names should not be entered in numeric characters.
	In the national code section, validation is required in order to enter valid national codes.
	The entry of non-numerical characters should be prevented as a patient number.
	There is a problem with the measurement turn: it is better to be entered by the system.
	It is better to set proper labels for each axis of the charts.
	Fonts in green are not good at all.
	Submit button is pale and blurred.
	Numeric default values have been defined in blank fields, while it is better to enter dashes if no values are entered.
	Mandatory fields are required to be marked with an asterisk.
Endocrinologists	Instead of insulin-dependent diabetes, type 1 diabetes must be used.
	In the diabetes treatment section, the term “others” should be deleted.
	In the comorbidities section, the term “cataract” should be deleted.
	In the neuropathy section, the term “behavioral disorders” should be deleted.
	In eye diseases, the term “glaucoma” should be added.
	The term “goiter” should be written in the form of “simple goiter.”
	The term “intermittent claudication” should be written instead of “ischemic pain of organs” and should be placed in the section of cardiovascular diseases.
	The time of blood glucose measurement needs to be determined.
Type 2 diabetic patients	Home icon should be used next to the term “homepage.”
	Blood glucose list numbers should be displayed.
	Instead of millimeters of Mercury, the unit of centimeters of Mercury should be used for hypertension.
	Abnormalities in the graph should be shown with a different color.
	The patient is required to read the guide.

XSL•FO
RenderX

Table 5. The demographic characteristics and distribution difference of participants in control and intervention groups.

Variable	Frequency (percent)		
	Control group (n=27)	Intervention group (n=26)	*P* value of distribution difference of variables in 2 groups
Gender			
Male	11 (41)	15 (58)	.28
Female	16 (59)	11 (42)	
Age group (year)			
≤30	1 (4)	0	.53
30-50	5 (18)	9 (35)	
≤50	21 (78)	17 (65)	
Marital status			
Single	3 (11)	0	.24
Married	24 (89)	26 (100)	
Employment status			
Employed	12 (44)	18 (69)	.01
Unemployed	15 (56)	8 (31)	
Education level			
Diploma	7 (26)	5 (19)	.38
Associate and BSc	17 (63)	16 (62)	
MSc and PhD	3 (11)	5 (19)	
Family history of Diabetes Mellitus			
Yes	21 (78)	20 (77)	>.99
No	6 (22)	6 (23)	
Type of drug taken			
Insulin	9 (33)	5 (19)	.85
Oral	11 (41)	17 (66)	
Insulin and oral	7 (26)	4 (15)	
History of high blood pressure			
Yes	17 (63)	15 (58)	.78
No	10 (37)	11 (42)	
Access to measurement tools at home			
Glucometer	6 (22)	3 (12)	.53
Glucometer, sphygmomanometer, scale	10 (37)	16 (61)	
Glucometer, sphygmomanometer	9 (33)	3 (12)	
Glucometer, scale	2 (8)	4 (15)	

Usability Evaluation of Diabetic Personal Health Records App

To have a preliminary usability evaluation of DPHR app, the usability questionnaire was submitted to 20 PhD and 30 MSc students of medical informatics. Of them, 11 PhD and 8 MSc students responded. In addition, 1 medical informatics expert, 2 endocrinologists, and 6 type 2 diabetic patients contributed in this respect. The usability evaluation also lasted for almost 50 days.

To evaluate the usability of Web-based DPHR app by diabetic patients, 6 people, including 3 women and 3 men, participated in this study. In terms of level of education, 4 participants had BSc and 2 had high school diploma degrees. The mean age of the patients was 50 years, and the average time of evaluation sessions was 30 minutes. The details related to the demographic profile and the opinions of diabetic patients and experts participating in usability evaluation of DPHR app are indicated in Tables 3 and 4.

Descriptive Analysis

The descriptive analysis of both intervention and control groups was conducted separately using frequency, percent, and mean for qualitative and quantitative variables. In the control group, there were 16 out of 36 women (59%) and 11 men (41%) with a mean age of 57 years. In terms of level of education, 17 out of 36 individuals (63%) held associate and BSc degrees. Considering employment status, 12 out of 36 participants (44%) were employed and 15 participants (56%) were unemployed. On an average, they worked with a computer for 8.5 hours per week.

In the intervention group, there were 11 out of 36 women (42%) and 15 men (58%) with a mean age of 52 years. In terms of level of education, 16 out of 36 individuals (62%) held associate and BSc degrees. Considering employment status, 18 out of 36 participants (69%) were employed. On average, they worked with a computer for 18 hours per week. Details relating to the demographic characteristics and distribution difference of participants in control and intervention groups are presented in Table 5.

Confounder Analysis

To analyze the equality of distribution for some variables which were likely to be confounder variables, the chi-square test was applied to qualitative variables, including gender, age group (year), marital status, employment status, education level, family history of DM, types of drug taken, history of high blood pressure, access to measurement tools at home, computer literacy, range of working time with a computer (hour), and range of disease length (year). The test results demonstrated no significant differences in the distribution of the variables between the intervention and control groups other than the variable of the range of working time with a computer, where participants in the intervention group at the baseline stage had spent more time working with a computer (Table 4).

Normality Analysis

To compare self-care indicators and their dimensions in the control and intervention groups, their normality was first evaluated using the Kolmogorov-Smirnov (K-S) test, which revealed that the distribution of self-care indicators was normal. Moreover, in the control group at the baseline stage, the dimensions of self-care status, including information of normal values, information of change trend, information of the latest measurement values, and information of training tips were normal in terms of distribution; however, other dimensions such as information of the physician’s advices, visit information, and information of date and time of the latest measurement values were abnormal.

Additionally, in the intervention group, the dimensions of self-care status including information of normal values, information of change trend, information of the latest measurement values, information of date and time of the latest measurement values, and information of training tips were normal in terms of distribution; however, other dimensions such as information of the physician’s advice and visit information were abnormal.

Inferential Analysis

In continuation, the parametric tests such as independent T-test were employed for analyzing the distribution of self-care indicators and their dimensions with normal distribution in both groups at the baseline stage, and for dimensions with abnormal distribution, the nonparametric tests such as Mann-Whitney U test was applied. The test results indicated that distribution of self-care indicators and their dimensions other than the sixth dimension, that is, information of training tips, were not significantly different in both groups at the baseline stage.

To compare the scores of self-care indicators in both groups, the independent T-test was employed. Test results revealed that there was a significant difference in terms of self-care indicators in both groups of diabetic patients.

Moreover, we found a significant difference in the dimensions of self-care indicators, including information of normal values, information of the trend of change, information of the latest measurement values, and information of date and time of the latest measurement values. However, no difference was observed in other dimensions such as information of the physician’s advice, visit information, and information of the training tips.

In addition, the independent T-test was utilized to compare the scores of weight, HbA1c, serum creatinine, HDL, LDL, total cholesterol, and triglyceride in control and intervention groups. The test results revealed no significant difference between any of them. Details relating to the comparison of average difference of self-care indicator, its dimensions and clinical outcomes in control and intervention groups are outlined in Table 6.

XSL•FO
RenderX

Table 6. A comparison of the average difference of self-care status, its dimensions, and clinical outcomes in control and intervention groups.

Outcome measure	Dimensions	Mean (SD)			
		Control group (n=27)	Intervention group (n=26)	*P* value	95% CI
Self-care status	Information of normal values	1 (1)	2.8 (1)	<.001	(-2.3 to -1.1)
	Information of change trend	-0.2 (0.8)	1.3 (2)	<.001	(-2.4 to -0.6)
	Information of physicians' advice	0.1 (0.4)	0.2 (0.4)	.73	(-0.2 to 0.2)
	Visit information	0.26 (0.447)	0.3 (0.5)	.94	(-0.2 to 0.2)
	Information of latest measurement values	0.04 (1)	1.9 (1.6)	<.001	(-2.6 to -1.1)
	Information of date and time of latest measurement values	-0.2 (1)	2 (1.5)	<.001	(-2.9 to -1.4)
	Information of training tips	1.7 (1)	2 (2.6)	.51	(-1.4 to 0.7)
	Self-care indicator	2.8 (2.4)	10.6 (4.5)	<.001	(-9.7 to -5.8)
Clinical outcomes	Weight	0.03 (1)	-0.9 (2.4)	.08	(-0.1 to 2.0)
	HbA1c	0.2 (0.1)	-0.2 (0.1)	.22	(-0.2 to 0.8)
	HDL	-0.4 (11)	-5 (17)	.298	(-4.6 to 14.6)
	LDL	4 (35)	-3 (23)	.44	(-10.8 to 24.8)
	Total cholesterol	-14 (57)	-6 (26)	.75	(-64.3 to 47.8)
	Triglyceride	-4.5 (157)	-26 (45)	.56	(-53.6 to 96.9)
	Serum creatinine	0.05 (0.3)	-0.01 (0.3)	.42	(-0.1 to 0.2)

Table 7. The comparison of visit adherence in control and intervention groups.

Visit adherence	Group		Total	*P* value
	Control	Intervention		
No	3	1	4	.61
Yes	24	25	49	
Total	27	26	53	

Fisher's exact test was also used to compare visit adherence scores in both groups, and the test results showed no significant difference. Table 7 provides the comparison of visit adherence in control and intervention groups.

The change trend of mean for FBS, 2 hours after lunch and 2 hours after dinner, indicates fluctuations in the 2-hour-after-dinner trend. However, the variations during the 2 hours after lunch had a steady state until the fifth measurement and then they had an unstable mode. Furthermore, FBS was almost in an unstable mode, although, a reducing pattern was observed in the final measurements. Figure 2 provides the change trend of mean related to blood sugar in intervention group.

The mean trend of systolic and diastolic blood pressure demonstrates the steady and stable trends, and no increasing or decreasing patterns were observed. Figure 3 provides the change trend of mean related to blood pressure in intervention group.

The relationship between covariates and self-care indicators for the 2-state variables such as gender and employment status was analyzed through T-test. Furthermore, the 1-way analysis of variance (ANOVA) was applied to multi-state variables such as education level, age group, length of disease, computer literacy, and the range of working time with a computer. The test results showed a significant difference in the variable of employment status in a way that the individuals employed obtained higher scores than the unemployed ones. However, there were no significance differences between other aforementioned covariates and self-care indicators.

XSL•FO
RenderX

Figure 2. Change trend of mean related to blood sugar in intervention group. BS: blood sugar; FBS: fasting blood sugar.

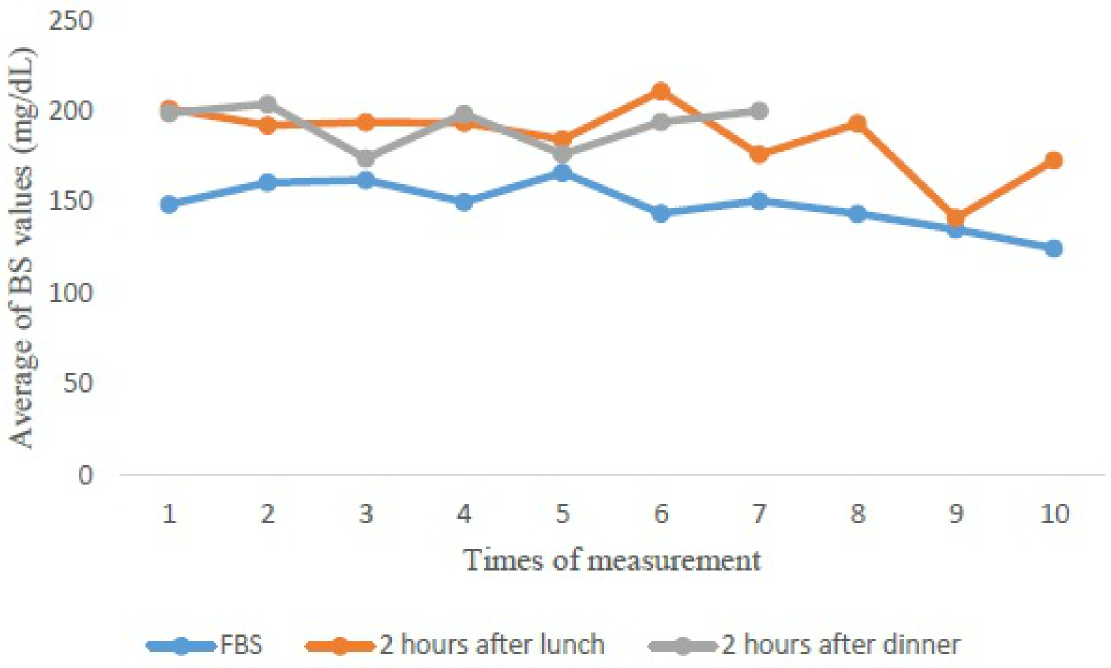
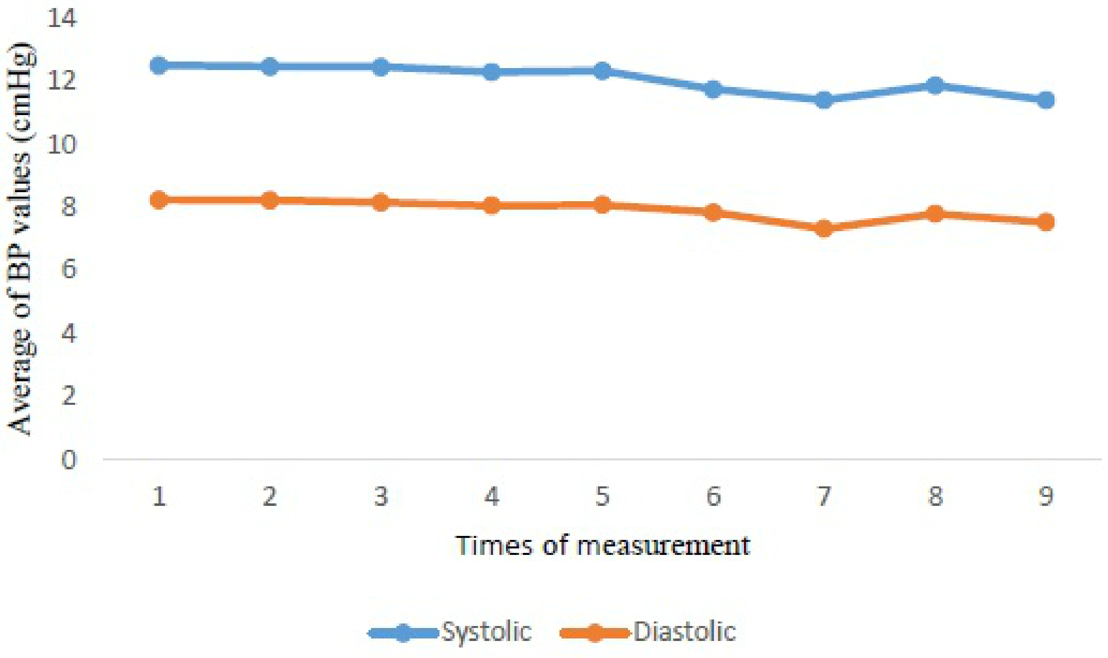


Figure 3. Change trend of mean related to blood pressure in intervention group. BP: blood pressure.



Discussion

Principal Findings

Impact of Diabetic Personal Health Records on Primary Outcome Measures

The test results demonstrated that the Web-based DPHR app had a positive impact on the primary outcome measure, namely in the status of self-care in general, and in 4 of its dimensions, in particular, including information of normal values, information of change trend, information of the latest measurement values, and information of date and time of the latest measurement values. Investigations show few studies on the relationship between PHR and self-care status. It is pointed out that in studies available in the field, the self-care index has been usually defined relatively homogeneous, but the important point is that any self-care index studied by the researchers may have different dimensions and questions, which can affect the efficiency of intervention. The self-care index assessed by the researchers in this study had 7 main dimensions and a total of 56 questions.

XSL•FO
RenderX

Impact of Diabetic Personal Health Records on Secondary Outcome Measures

Researchers have not found any positive effect between Web-based DPHR app and clinical outcomes including weight, HbA1c, serum creatinine, HDL, LDL, total cholesterol, and triglyceride. There are differences and sometimes contradictions among existing studies for the effect of PHR interventions on clinical outcomes associated with diabetics such as HbA1c and lipid profile. Most studies refer to a positive relationship of PHR [11,30,47,48], though several studies indicated no positive effect [30,49]. Researchers revealed that such contradictions could have several reasons, the most important ones being how to design the app, type of study design, duration of study, and study attrition rate. It is important to note that the impact of some of the positive studies is in doubt because of limitations and bias. Moderate to high risk of bias has been reported in 4 studies assessing the interventions [50]. It seems that relatively short-term duration of the study is the primary reason for the lack of positive association between DPHR and clinical outcomes in this study. Conducting a systematic review on the effectiveness of DPHR and clinical outcomes in patients with type 2 diabetes can be very useful.

Usability Evaluation of Diabetic Personal Health Records App

In this study, usability-testing process was carried out through a scientific process and with the participation of 20 specialists in medical informatics, 2 endocrinologists and 6 diabetics. The proper sample size is essential for usability testing, so that research, which found that up to 80% of usability has issues, can be determined with 5 to 8 participants [51]. Based on our usability testing, Jakob Nielsen's general principles for interaction design such as error prevention, consistency and standards, aesthetic and minimalist design, recognition rather than recall, and help and documentation are required to be considered, and which were addressed on our app through iterative refinement process [52]. One of the most important cases in intervention implementation is usability testing [53] but this is often neglected, with up to 60% of diabetes-related websites having a minimum of 4 usability errors. So it can be said that the gold standard for intervention development should be represented to obtain high intervention effectiveness [54].

Impact of Diabetic Personal Health Records on Visit Adherence

The results of Fisher’s exact test on the comparison of visit adherence scores revealed no significant difference in both intervention and control groups. Given the importance of disease follow-up and the treatment of diabetic patients by attending physicians, the patients paid attention to their visits, and their efforts in terms of planned visit adherence were implicit.

Impact of Diabetic Personal Health Records on Blood Sugar

The change trend of mean for FBS 2 hours after lunch and 2 hours after dinner exhibited that there was generally a sinus trend in all the above-mentioned cases. The main difference between this study and those in the related literature is that in our study we measured trends in blood sugar levels for 10 times, whereas such values were usually compared before and after intervention in most studies. In a study by Davies et al, by using DPHR for 6 months, blood sugar levels had improved in both groups, particularly in the intervention group [47].

Impact of Diabetic Personal Health Records on Blood Pressure

The change trend of mean for systolic and diastolic blood pressure in the intervention group revealed steady and stable trends, and no increasing or decreasing patterns were observed. Previous studies indicated no difference was noticed in the improvement of blood pressure in control and intervention groups [30,49]. No improvement was also found in levels of systolic blood pressure values in a study conducted by Dijkstra and only slight improvements were observed in diastolic values [55].

Correlation Between Diabetic Personal Health Records App and Covariates

The results of the chi-square test and ANOVA for analyzing the correlation between the trial covariates and the self-care index showed that there was a significant difference in the variable of employment status in a way that the individuals employed obtained higher scores than the unemployed ones. Moreover, in comparison with the previous studies, there was no significant difference in self-care index scores among type 2 diabetic patients in terms of their marital status, because more than 90% of the participants in this study were married. In a study by Bohanny et al, the scores of self-care behaviors were significantly higher in married individuals than those obtained by single people [56]. However, such conflicting results require more investigations.

Strengths of Study

The strengths of our study are as follows: the evidence-based development process of the DPHR app (based on a systematic review), the inclusion of local experts’ opinions, and the iterative refinement of the app using the usability techniques. Both the value of evidence-based content development and the importance of usability testing in the app development process have been emphasized in several studies [53,54,57], pointing to the fact that such considerations have rarely been used in similar works [54].

Limitations

A few limitations of our trial are as follows:

1. The primary need to recruit participants with minimum computer skills and Internet literacy was a limiting factor. Generally, in Web-based interventions, issues such as digital divide, computer literacy, age, and interest in technology can be effective in participant recruitment. The young, computer literates, and those having access to the Internet usually have a strong tendency toward participating in such studies. This trial is not an exception in principle. Such tendencies may present bias in our findings, and thus, our trial may not necessarily represent the actual distribution of the population being studied.

XSL•FO
RenderX

2. The trial sample only represents type 2 diabetes patients. It is possible that the findings could not be generalized to other types of diabetes disease.

3. Due to the nature of intervention, the investigators frequently requested the presence of the participants for the interviews. This induced discomfort to some of the participants. To address such issues, we provided financial incentives such as free visits and laboratory testing in order to encourage better involvement.

4. The other limitation of our study was the passive participation of some patients during the study, especially at the early stages. Therefore, lack of participation could lead to loss of useful information from patients. Therefore, reminders via telephone contact and SMS text messaging were employed.

5. Considering the limitations of our patient sample, the results should be interpreted cautiously [58]. The small sample size in our research may affect the representativeness and generalizability of the findings [59].

6. We concentrated on the comparison of primary and secondary outcome measures in our RCT study design, indicating explicitly a variation in consumer self-care status regarding the complexity of intervention [60].

Implications and Future Directions

The methods and findings of this study are expected to be used as a suitable platform for other endocrine and metabolic disorders as well as other fields of medical science studies to assess the impact of the PHR intervention on the self-care index and clinical outcomes. Multi-center study is proposed to be carried out on a broader level to ensure the effectiveness of the Web-based DPHR intervention on the self-care index and clinical outcomes. In this case, endocrinologists and patients with type 2 diabetes will be involved in the study in a wider range, which can be very important in the generalizability of the study findings, especially for developing countries. Moreover, the impact of DPHR efficiency on the level of decision-making of the endocrinologists is recommended to be evaluated through a proper RCT study.

Diabetes knowledge, involvement by health care providers, patient empowerment, and enhanced health care status will be promoted hopefully by the improved self-care status recommended by the Web-based DPHR to patients with type 2 diabetes mellitus.

Conclusions

As a result of a systematic review of literature together with the representative sample of endocrinologists in Iran, a consensus was achieved on a Web-based DPHR model to improve self-care for type 2 diabetic patients. However, to take advantages of the DPHR, the given Web-based DPHR app was implemented and evaluated on type 2 diabetic patients after iterative refinement of the app user interface, using usability techniques. We found as a result of the Web-based DPHR app that the self-care scores in the intervention group were significantly higher than those of the control group. In total, we found no correlation between the DPHR app and covariates, including planned visit adherence, HbA1c, serum creatinine, HDL, LDL, total cholesterol, weight, the change trend of the mean blood glucose, and blood pressure.

Acknowledgments

This trial is the first author’s PhD dissertation which has been supported by a grant [grant # 921835] from Mashhad University of Medical Sciences Research Council. The authors would like to thank Dr. Hamid-Reza Dehghan and Dr. Hadi Ghasemi for their contributions to Web-based DPHR app development, as well as diabetes specialists from Mashhad Endocrine Research Center for their contribution in Delphi Technique.

Authors' Contributions

MT is the principal investigator and conceived the trial. He was responsible for overall administration of the grant. MT and AA were primarily responsible for development of the DPHR app. ZM and RA assisted in trial coordination. MA and HT Provided expertise in the RCT design and analysis. All authors participated in the critical revision and protocol design.

Conflicts of Interest

None declared.

References

1. World Health Organization. Geneva Noncommunicable diseases URL: http://www.who.int/mediacentre/factsheets/fs355/en/ [accessed 2015-05-26] [WebCite Cache ID 6czA1rqGT]
2. Tani S, Marukami T, Matsuda A, Shindo A, Takemoto K, Inada H. Development of a health management support system for patients with diabetes mellitus at home. J Med Syst 2010 Jun;34(3):223-228. [Medline: 20503606]
3. Liang X, Wang Q, Yang X, Cao J, Chen J, Mo X, et al. Effect of mobile phone intervention for diabetes on glycaemic control: a meta-analysis. Diabet Med 2011 Apr;28(4):455-463. [doi: 10.1111/j.1464-5491.2010.03180.x] [Medline: 21392066]
4. Lim S, Kim S, Kim JI, Kwon MK, Min SJ, Yoo SY, et al. A Survey on Ubiquitous Healthcare Service Demand among Diabetic Patients. Diabetes Metab J 2011 Feb;35(1):50-57 [FREE Full text] [doi: 10.4093/dmj.2011.35.1.50] [Medline: 21537413]

XSL•FO
RenderX

5. Lyles CR, Harris LT, Le T, Flowers J, Tufano J, Britt D, et al. Qualitative evaluation of a mobile phone and web-based collaborative care intervention for patients with type 2 diabetes. Diabetes Technol Ther 2011 May;13(5):563-569. [doi: 10.1089/dia.2010.0200] [Medline: 21406018]
6. Alberti KG, Zimmet PZ. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. Diabet Med 1998 Jul;15(7):539-553. [doi: 10.1002/(SICI)1096-9136(199807)15:7<539::AID-DIA668>3.0.CO;2-S] [Medline: 9686693]
7. Chodosh J, Morton SC, Mojica W, Maglione M, Suttorp MJ, Hilton L, et al. Meta-analysis: chronic disease self-management programs for older adults. Ann Intern Med 2005 Sep 20;143(6):427-438. [Medline: 16172441]
8. Ellis S, Speroff T, Dittus R, Brown A, Pichert J, Elasy T. Diabetes patient education: a meta-analysis and meta-regression. Patient Educ Couns 2004 Jan;52(1):97-105. [Medline: 14729296]
9. Norris SL, Nichols PJ, Caspersen CJ, Glasgow RE, Engelgau MM, Jack L, et al. Increasing diabetes self-management education in community settings. A systematic review. Am J Prev Med 2002 May;22(4 Suppl):39-66. [Medline: 11985934]
10. Rizvi AA, Sanders MB. Assessment and monitoring of glycemic control in primary diabetes care: monitoring techniques, record keeping, meter downloads, tests of average glycemia, and point-of-care evaluation. J Am Acad Nurse Pract 2006 Jan;18(1):11-21. [doi: 10.1111/j.1745-7599.2006.00092.x] [Medline: 16403208]
11. Holbrook A, Thabane L, Keshavjee K, Dolovich L, Bernstein B, Chan D, et al. Individualized electronic decision support and reminders to improve diabetes care in the community: COMPETE II randomized trial. CMAJ 2009 Jul 7;181(1-2):37-44 [FREE Full text] [doi: 10.1503/cmaj.081272] [Medline: 19581618]
12. Hogan P, Dall T, Nikolov P, American Diabetes Association. Economic costs of diabetes in the US in 2002. Diabetes Care 2003 Mar;26(3):917-932. [Medline: 12610059]
13. Funnell MM, Brown TL, Childs BP, Haas LB, Hosey GM, Jensen B, et al. National standards for diabetes self-management education. Diabetes Care 2010 Jan;33 Suppl 1:S89-S96 [FREE Full text] [doi: 10.2337/dc10-S089] [Medline: 20042780]
14. Kickbusch I. Self-care in health promotion. Soc Sci Med 1989;29(2):125-130. [Medline: 2665107]
15. Alavinia M, Ghotbi M, Mahdavi-Hazaveh A, Kermanchi J, Nasli-Esfahani A, Yarahmadi S. Sums. Tehran: Sepidbarg Publishing; 2012. National program to prevent and control type 2 diabetes [In Persian] URL: http://fhc.sums.ac.ir/files/gh-vagir/ayin%20name/diabetes_guideline_91_final_-_Copy.pdf [accessed 2016-10-10] [WebCite Cache ID 6l9FWIFxq]
16. Rakhshanderou S, Heidarnia A, Rajab A. The Effect of Health Education on Quality of Life in Diabetic Patients Referring to Iran Diabetes Association [In Persian]. Daneshvar Med 2006;13(5):15-20.
17. International Diabetes Federation. Diabetes in Iran 2014 and Iran vs world prevalence of diabetes URL: https://www.idf.org/membership/mena/iran [accessed 2015-11-13] [WebCite Cache ID 6czMbAV8p]
18. World Health Organization. Global Status Report on Noncommunicable Diseases 2014. Geneva: World Health Organization; 2014:1-298.
19. Benhamou P. Improving diabetes management with electronic health records and patients' health records. Diabetes Metab 2011 Dec;37(Suppl 4):S53-S56 [FREE Full text] [doi: 10.1016/S1262-3636(11)70966-1] [Medline: 22208711]
20. Kim M, Johnson K. Personal health records: evaluation of functionality and utility. J Am Med Inform Assoc 2002;9(2):171-180 [FREE Full text] [Medline: 11861632]
21. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 2006;13(2):121-126 [FREE Full text] [doi: 10.1197/jamia.M2025] [Medline: 16357345]
22. Archer N, Fevrier-Thomas U, Lokker C, McKibbon K, Straus S. Personal health records: a scoping review. J Am Med Inform Assoc 2011;18(4):515-522 [FREE Full text] [doi: 10.1136/amiajnl-2011-000105] [Medline: 21672914]
23. Tenforde M, Jain A, Hickner J. The value of personal health records for chronic disease management: what do we know? Fam Med 2011 May;43(5):351-354 [FREE Full text] [Medline: 21557106]
24. Nokes KM, Hughes V, Santos R, Bang H. Creating a paper-based personal health record for HIV-infected persons. J Assoc Nurses AIDS Care 2012;23(6):539-547 [FREE Full text] [doi: 10.1016/j.jana.2011.11.004] [Medline: 22512926]
25. Wagholikar A, Fung M, Nelson C. Improving self-care of patients with chronic disease using online personal health record. Australas Med J 2012;5(9):517-521 [FREE Full text] [doi: 10.4066/AMJ.2012.1358] [Medline: 23115588]
26. Tenforde M, Nowacki A, Jain A, Hickner J. The association between personal health record use and diabetes quality measures. J Gen Intern Med 2012 Apr;27(4):420-424 [FREE Full text] [doi: 10.1007/s11606-011-1889-0] [Medline: 22005937]
27. Bond G, Burr R, Wolf F, Price M, McCurry S, Teri L. The effects of a web-based intervention on the physical outcomes associated with diabetes among adults age 60 and older: a randomized trial. Diabetes technology & therapeutics 2007;9(1):52-59. [doi: 10.1089/dia.2006.0057]
28. Shea S, Weinstock R, Teresi J, Palmas W, Starren J, Cimino J, IDEATel Consortium. A randomized trial comparing telemedicine case management with usual care in older, ethnically diverse, medically underserved patients with diabetes mellitus: 5 year results of the IDEATel study. J Am Med Inform Assoc 2009;16(4):446-456 [FREE Full text] [doi: 10.1197/jamia.M3157] [Medline: 19390093]

XSL•FO
RenderX

29. McCarrier K, Ralston J, Hirsch I, Lewis G, Martin D, Zimmerman F, et al. Web-based collaborative care for type 1 diabetes: a pilot randomized trial. Diabetes Technol Ther 2009 Apr;11(4):211-217 [FREE Full text] [doi: 10.1089/dia.2008.0063] [Medline: 19344195]
30. Ralston J, Hirsch I, Hoath J, Mullen M, Cheadle A, Goldberg H. Web-based collaborative care for type 2 diabetes: a pilot randomized trial. Diabetes Care 2009 Feb;32(2):234-239 [FREE Full text] [doi: 10.2337/dc08-1220] [Medline: 19017773]
31. Brown LL, Lustria ML, Rankins J. A review of web-assisted interventions for diabetes management: maximizing the potential for improving health outcomes. J Diabetes Sci Technol 2007 Nov;1(6):892-902 [FREE Full text] [Medline: 19885163]
32. McMahon GT, Gomes HE, Hickson HS, Hu TM, Levine BA, Conlin PR. Web-based care management in patients with poorly controlled diabetes. Diabetes Care 2005 Jul;28(7):1624-1629 [FREE Full text] [Medline: 15983311]
33. Tara M, Mazloum-Khorasani Z, Babaee M. Mashhad, Iran: Mashhad University of Medical Sciences; 2014. Design and Evaluation of a Home Guideline-based Decision Support Model to Improve Self-Management in Patients with Diabetes URL: http://research.mums.ac.ir/webdocument/load.action?webdocument_code=5000&masterCode=8106499 [accessed 2016-09-30] [WebCite Cache ID 6kuVCZ5Cs]
34. Hieftje K, Edelman E, Camenga D, Fiellin L. Electronic media-based health interventions promoting behavior change in youth: a systematic review. JAMA Pediatr 2013 Jun;167(6):574-580 [FREE Full text] [doi: 10.1001/jamapediatrics.2013.1095] [Medline: 23568703]
35. Pal K, Eastwood S, Michie S, Farmer A, Barnard M, Peacock R, et al. Computer-based diabetes self-management interventions for adults with type 2 diabetes mellitus. Cochrane Database Syst Rev 2013(3):Cd008776. [doi: 10.1002/14651858.CD008776.pub2] [Medline: 23543567]
36. Azizi A, Aboutorabi R, Mazloum-Khorasani Z, Hoseini B, Tara M. Diabetic Personal Health Record: A Systematic Review Article. Iranian Journal of Public Health 2016 (forthcoming).
37. Nielsen J, Molich R. Heuristic evaluation of user interfaces. : ACM; 1990 May 01 Presented at: Conference on Human Factors in Computing Systems; 1990-05-01; Seattle, Washington, USA p. 249-256 URL: http://dl.acm.org/citation.cfm?id=97243&picked=prox
38. Diabetic Association of Khorasan Razavi [In Persian]. Statistics of diabetic patients in Mashhad URL: http://diabet.blogsky.com/ [accessed 2015-11-13] [WebCite Cache ID 6czOk9tZB]
39. Rubin RR, Peyrot M. Quality of life and diabetes. Diabetes Metab Res Rev 1999;15(3):205-218. [Medline: 10441043]
40. Saeidpour J, Jafari M, Ghazi-Asgar M, Dayani-Dardashti H. Effect of educational program on quality of life in diabetic patients [In Persian]. Journal of Health Administration 2013;16(52):26-36.
41. Toobert DJ, Hampson SE, Glasgow RE. The summary of diabetes self-care activities measure: results from 7 studies and a revised scale. Diabetes Care 2000 Jul;23(7):943-950 [FREE Full text] [Medline: 10895844]
42. Bakeridi. Diabetes Knowledge Questionnaire URL: https://www.bakeridi.edu.au/Assets/Files/AusDiab_Diabetes_Knowledge_Questionnaire_04_05.pdf [accessed 2016-09-01] [WebCite Cache ID 6kuXo29DB]
43. Michigan Diabetes Research and Training Centers. Diabetes knowledge test URL: http://diabetesresearch.med.umich.edu/Tools_SurveyInstruments.php [accessed 2016-09-30] [WebCite Cache ID 6kuY4DDc8]
44. Diabetes Initiative. Patient's Diabetes Knowledge Questionnaire URL: http://www.diabetesinitiative.org/resources/tools/documents/8-GATE-KNOWLEDGEQUESTIONAIRE_web.pdf [accessed 2016-09-01] [WebCite Cache ID 6kuYGbffF]
45. Yu C, Parsons J, Mamdani M, Lebovic G, Hall S, Newton D, et al. A web-based intervention to support self-management of patients with type 2 diabetes mellitus: effect on self-efficacy, self-care and diabetes distress. BMC Med Inform Decis Mak 2014;14:117 [FREE Full text] [doi: 10.1186/s12911-014-0117-3] [Medline: 25495847]
46. Azizi A, Aboutorabi R, Mazloum-Khorasani Z, Afzal-Aghaea M, Tara M. Development, Validation, and Evaluation of Web-Based Iranian Diabetic Personal Health Record: Rationale for and Protocol of a Randomized Controlled Trial. JMIR Res Protoc 2016;5(1):e39 [FREE Full text] [doi: 10.2196/resprot.5201] [Medline: 26964572]
47. Davies M, Quinn M. Patient-held diabetes record promotes seamless shared care. Guidelines in Practice 2001;4(11) [FREE Full text]
48. Aghamolaei T, Eftekhar H, Mohammad K, Sobhani A, Shojaeizadeh D, Nakhjavani M. Influence of Educational Intervention Using Interaction Approach on Behavior Change, Hemoglobin A1c and Health-Related Quality of Life in Diabetic Patients (In Persian). Journal of School of Public Health and Institute of Public Health Research 2005;3(4):1-2 [FREE Full text]
49. Grant R, Wald J, Schnipper J, Gandhi T, Poon E, Orav E, et al. Practice-linked online personal health records for type 2 diabetes mellitus: a randomized controlled trial. Arch Intern Med 2008 Sep 8;168(16):1776-1782 [FREE Full text] [doi: 10.1001/archinte.168.16.1776] [Medline: 18779465]
50. Ko H, Turner T, Jones C, Hill C. Patient-held medical records for patients with chronic disease: a systematic review. Quality & safety in health care 2010;19(5):e41. [doi: 10.1136/qshc.2009.037531] [Medline: 20511601]
51. Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. J Biomed Inform 2004 Feb;37(1):56-76 [FREE Full text] [doi: 10.1016/j.jbi.2004.01.003] [Medline: 15016386]
52. Yu CH, Parsons JA, Hall S, Newton D, Jovicic A, Lottridge D, et al. User-centered design of a web-based self-management site for individuals with type 2 diabetes – providing a sense of control and community. BMC Med Inform Decis Mak 2014 Jul 23;14(1):1-15. [doi: 10.1186/1472-6947-14-60] [Medline: 25056379]

XSL•FO
RenderX

53. Fu L, Salvendy G. The contribution of apparent and inherent usability to a user's satisfaction in a searching and browsing task on the Web. Ergonomics 2002 May 15;45(6):415-424. [doi: 10.1080/00140130110120033] [Medline: 12061966]
54. Yu CH, Bahniwal R, Laupacis A, Leung E, Orr MS, Straus SE. Systematic review and evaluation of web-accessible tools for management of diabetes and related cardiovascular risk factors by patients and healthcare providers. J Am Med Inform Assoc 2012;19(4):514-522 [FREE Full text] [doi: 10.1136/amiajnl-2011-000307] [Medline: 22215057]
55. Dijkstra RF, Braspenning JC, Huijsmans Z, Akkermans RP, van Ballegooie E, ten Have P, et al. Introduction of diabetes passports involving both patients and professionals to improve hospital outpatient diabetes care. Diabetes Res Clin Pract 2005 May;68(2):126-134. [doi: 10.1016/j.diabres.2004.09.020] [Medline: 15860240]
56. Bohanny W, Wu SV, Liu C, Yeh S, Tsay S, Wang T. Health literacy, self-efficacy, and self-care behaviors in patients with type 2 diabetes mellitus. J Am Assoc Nurse Pract 2013 Sep;25(9):495-502. [doi: 10.1111/1745-7599.12017] [Medline: 24170654]
57. Seidman JJ, Steinwachs D, Rubin HR. Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites. J Med Internet Res 2003 Nov 27;5(4):e30 [FREE Full text] [doi: 10.2196/jmir.5.4.e30] [Medline: 14713658]
58. Greenhalgh T, Hinder S, Stramer K, Bratan T, Russell J. Adoption, non-adoption, and abandonment of a personal electronic health record: case study of HealthSpace. BMJ 2010 Nov 16;341(nov16 1):c5814. [doi: 10.1136/bmj.c5814]
59. Liu C, Tsai Y, Jang F. Patients' acceptance towards a web-based personal health record system: an empirical study in Taiwan. International journal of environmental research and public health 2013;10(10):5191-5208. [doi: 10.3390/ijerph10105191]
60. Arguel A, Lau AY, Dennis S, Liaw S, Coiera E. An internet intervention to improve asthma management: rationale and protocol of a randomized controlled trial. JMIR Res Protoc 2013;2(2):e28 [FREE Full text] [doi: 10.2196/resprot.2695] [Medline: 23942523]

Abbreviations

ANOVA: analysis of variance
BMI: body mass index
DM: diabetes mellitus
DPHR: diabetic personal health records
EHR: electronic health record
FBS: fasting blood sugar
HbA $_{1c}$: glycated hemoglobin
LDL: low-density lipoprotein
PHR: personal health record
RCT: randomized controlled trial
SMS: short message service
WHO: World Health Organization

Edited by G Eysenbach; submitted 02.08.16; peer-reviewed by D Ose, T Vaughan; comments to author 17.08.16; revised version received 05.09.16; accepted 07.09.16; published 21.10.16.

Please cite as:
Azizi A, Aboutorabi R, Mazloum-Khorasani Z, Afzal-Aghaea M, Tabesh H, Tara M
Evaluating the Effect of Web-Based Iranian Diabetic Personal Health Record App on Self-Care Status and Clinical Indicators: Randomized Controlled Trial
JMIR Med Inform 2016;4(4):e32
URL: http://medinform.jmir.org/2016/4/e32/
doi:10.2196/medinform.6433
PMID:27769953

©Amirabbas Azizi, Robab Aboutorabi, Zahra Mazloum-Khorasani, Monavar Afzal-Aghaea, Hamed Tabesh, Mahmood Tara. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 21.10.2016. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO RenderX

Review

A Review of Visual Representations of Physiologic Data

Rishikesan Kamaleswaran[1*], PhD; Carolyn McGregor[2*], PhD

[1]Center for Biomedical Informatics, Department of Pediatrics, University of Tennessee Health Science Center, Memphis, TN, United States

[2]University of Ontario Institute of Technology, Oshawa, ON, Canada

[*]all authors contributed equally

Corresponding Author:
Rishikesan Kamaleswaran, PhD
Center for Biomedical Informatics
Department of Pediatrics
University of Tennessee Health Science Center
50 N Dunlap St
Room 496R
Memphis, TN, 38111
United States
Phone: 1 4163002871
Fax: 1 4163002871
Email: rkamales@uthsc.edu

Abstract

Background: Physiological data is derived from electrodes attached directly to patients. Modern patient monitors are capable of sampling data at frequencies in the range of several million bits every hour. Hence the potential for cognitive threat arising from information overload and diminished situational awareness becomes increasingly relevant. A systematic review was conducted to identify novel visual representations of physiologic data that address cognitive, analytic, and monitoring requirements in critical care environments.

Objective: The aims of this review were to identify knowledge pertaining to (1) support for conveying event information via tri-event parameters; (2) identification of the use of visual variables across all physiologic representations; (3) aspects of effective design principles and methodology; (4) frequency of expert consultations; (5) support for user engagement and identifying heuristics for future developments.

Methods: A review was completed of papers published as of August 2016. Titles were first collected and analyzed using an inclusion criteria. Abstracts resulting from the first pass were then analyzed to produce a final set of full papers. Each full paper was passed through a data extraction form eliciting data for comparative analysis.

Results: In total, 39 full papers met all criteria and were selected for full review. Results revealed great diversity in visual representations of physiological data. Visual representations spanned 4 groups including tabular, graph-based, object-based, and metaphoric displays. The metaphoric display was the most popular (n=19), followed by waveform displays typical to the single-sensor-single-indicator paradigm (n=18), and finally object displays (n=9) that utilized spatiotemporal elements to highlight changes in physiologic status. Results obtained from experiments and evaluations suggest specifics related to the optimal use of visual variables, such as color, shape, size, and texture have not been fully understood. Relationships between outcomes and the users' involvement in the design process also require further investigation. A very limited subset of visual representations (n=3) support interactive functionality for basic analysis, while only one display allows the user to perform analysis including more than one patient.

Conclusions: Results from the review suggest positive outcomes when visual representations extend beyond the typical waveform displays; however, there remain numerous challenges. In particular, the challenge of extensibility limits their applicability to certain subsets or locations, challenge of interoperability limits its expressiveness beyond physiologic data, and finally the challenge of instantaneity limits the extent of interactive user engagement.

(JMIR Med Inform 2016;4(4):e31) doi:10.2196/medinform.5186

XSL•FO
RenderX

KEYWORDS

survey; human-centered computing; visualization application domains; information visualization; visualization systems and tools; visualization toolkits

Introduction

Two less formal reviews and one systematic review were published in the last decade, reporting positive impact of visual representations in the critical care setting. Sanderson et al provide a forward-looking analysis of representation of physiological data [1] in anesthesiology [2]. Drews and Westenskow review several graphical displays that facilitate rapid translation of physiological event knowledge for anesthesiologists [3]. An initial systematic review was published in 2007 by Görges and Staggers that reviews general physiologic data displays; however, with emphasis on surgical and anesthesiology specialities [4]. While those reviews provide important knowledge about the state of the art in physiologic data, they present only a partial aggregation of results, and limited knowledge that could be used to enhance the design of physiological visualizations. Furthermore, key elements such as the nature of visual variables utilized in the encoding, support for interactive exploration, and common design considerations were not discussed. All reviews focused on displays that support short-term patient monitoring tasks. Visualizations supporting longitudinal monitoring and interactive visual analysis of physiological data were not sufficiently addressed.

The aim of this specific review is of 3 parts: (1) identify the design decisions used in the development of novel physiologic visual representations; (2) review the utilization of temporal parameters namely: trajectory, frequency, and duration in visual designs using physiologic parameters; and (3) review the nature of interactive functions afforded for rich exploration tasks. With that in mind, this paper presents an analysis of a broad spectrum of physiological visual representations used at the bed-side, in the surgical ward, and for clinical research.

Methods

The review was conducted in 2 phases: the first phase identified the key terms to be included in the search strategy, while the second phase broadened the search strategy and used structured analysis method. In the first phase we used Google Scholar, and 25 papers were found to be relevant. The search was limited to the last 15 years and used a combination of keywords that were known to the author, such as "(physiologic* or clinical or hemodynamic) and (visual* or graphic*) and (interface or display)," where asterisk was used to search for terms that started with the specific key words. In the second phase, we used 6 prominent sources including: IEEE Explore, ACM Digital Library, MEDLINE, EMBASE, ISI Web of Science, and Google Scholar. A broad search strategy was used to capture as many representations as were possible. Index terms were used to filter articles and included "data display*," "diagnosis, computer-assisted," "monitoring, physiologic/methods*," "*computer graphics," "user-computer interface," "data display," "interview* or discussion* or questionnaire* or "focus group*" or qualitative or ethnograph* or fieldwork or "field work" or "key informant," "task performance and analysis," "graphic* adj2 display*".

For screening articles in the second phase, we used rigorous inclusion criteria (Textbox 1) that initially classified visualizations across 4 groups. The groups were (1) tabular displays, (2) waveform displays, (3) object displays, and (4) ecological displays. Inclusion criteria relating to outcome measures are divided into 3 sets of measures (Textbox 2). They include temporal and duration, human and qualitative factors, and quantitative measures. The physiological parameters tested are listed in Textbox 3. We placed a restriction in years from January 1, 1983 to August 1, 2016 and limited our results to human studies in critical care, anesthesiology, and surgery. We included snowballing of references and manual searches on Google Scholar and PubMed. This resulted in a total of 1262 titles generated for review. Relevant titles were identified using rigorous inclusion criteria (Textbox 1). In total, 171 titles were then designated for abstract review. Following that, 78 abstracts were selected for full review, and 39 papers were selected for inclusion in the analysis. Bias was mitigated by having 2 researchers screen independently, and differences were resolved through discussions until consensus was reached.

Textbox 1. Inclusion criteria.

Types of studies:

- Randomized controlled trials, cohort, case-control, and design studies.
- The review placed increasing preference for randomized control trials, followed by cohort, case-control, and finally design studies. Design studies are popular in the visualization community and were included to study results pertaining to user-evaluations.

Types of participants:

- Critical care nurses and physicians.
- Several studies have only tested interventions on physicians and excluded nurses, while other studies have used naive participants usually by recruiting undergraduates.

Types of interventions:

- Novel knowledge representations, numeric, waveform or metaphor-based displays.
- We focus on the intervention in which physiological display is not represented exclusively in waveform and/or static numerical forms.

XSL•FO
RenderX

Textbox 2. Reported metrics.

Temporal metrics:

- Time to detection of adverse event(s), time to diagnose(s), time to initiate treatment(s)

Human factors:

- NASA-TLX task load index score, satisfaction of intervention (Likert scales), number of participants, clinical expertise of participants, setting in which the trials were conducted, noise level of the environment, age of the participants, caffeine intake

Clinical relevance:

- Accuracy of diagnoses, accuracy of treatment

Textbox 3. Physiological parameters tested.

Physiological parameters:

- Central venous pressure (mm Hg)
- Mean left arterial pressure (mm Hg)
- Systemic vascular resistance
- ST segment depression of ECG (mm)
- Arterial oxygen saturation (%)
- Heart rate (bpm)
- Respiratory wave (impedance)
- End-tidal CO2
- Mean arterial blood pressure (mm Hg)
- Pulmonary vascular resistance
- Cardiac output (mL/min)
- Stroke volume (mL)
- Peripheral oxygen saturation (%)
- Respiratory rate (rpm)
- Pulse rate
- Mean pulmonary artery pressure (mm Hg)

Following the creation of the inclusion criteria, an online data extraction form was developed using Google forms and used to evaluate all papers. The data extraction form consisted of 6 sections that were identified as potential areas of interest. For each full paper reviewed, 74 questions were screened. Questions to be included in the data extraction forms were selected from themes identified in the pilot study. In particular, questions were generated to elicit detail about the study, design, and results from any human experiment or evaluations. Where appropriate the questions were marked as either not reported if data was missing, or not applicable if the question was a follow-up of a prior conditional question. The data was then thematically synthesized based on aggregations of results by descriptive codes. The thematic synthesis is presented using a series of matrices presented in the next section.

Results

Phase 1 and 2

All papers included in the analysis were passed through the data extraction form and resulted in an initial comprehensive matrix of results. Of 74 questions that were initially probed, results that yielded over 75% not reported, or not applicable across all papers analyzed were removed from our analysis. Then 39 variables were selected for inclusion in the initial matrix. Phase 1 results are summarized in the comprehensive matrix of design properties (Table 1 and Figure 1) and phase 2 results are summarized in the Comprehensive Matrix of Study Results (Table 2 and Table 3). The comprehensive matrix of design properties presents 10 variables which are divided between 2 tables. Variables appearing in (Table 1) are “Target Users”, “Year”, “Clinical Context”, “Number of Variables”, and “Display Type”.

“Target Users” relates to the clinical specialty, and “Year” is the approximate date the prototype was tested. Due to the difference between the dates of publication and evaluation, this value was approximated based on the date of submission of the article. “Clinical Context” conveys the copresence of contextual clinical information, and “Number of Variables” refers to the total number of physiological or clinical variables that were visible in a single screen. “Display Type” lists the types of graphics utilized by the paper belonging to one of: tabular, object, or metaphoric displays. “Color(s) Used” identifies the hue where available. “Pre-attentive Processing” lists particular

XSL•FO
RenderX

visual variables that were used in the visual representation such as: shape, size, and dimension. “Gestalts” refers to the designer’s use of grouping laws identified by Gestalt’s laws of perception: the use of proximity, similarity, closure, symmetry, and continuity as a means of discerning visual objects presented in the display [5]. Finally “Interactive Controls” refers to the ability for the display to support direct manipulation of one or more properties and “Iterative Design” identifies displays that were built using user-centered design approaches that include users into key decision making processes prior to the development of the display.

A second matrix, titled the Comprehensive Matrix of Study Results presents additional 11 variables that were identified in papers which presented study results. Table 2 lists “Setting,” “Study Type,” “Results Reported,” “Realism,” “Cognitive Workload,” “Historic Trends,” “Visual Encoding for Temporal Trajectory,” “Visual Encoding for Duration,” “Visual Encoding for Frequency”. Table 3 lists “Counter-balanced for Display or Scenario,” “Were Scenarios Clinically Relevant,” and “Function Supporting Case-controlled Analysis.” “Setting” describes the location where the study was physically conducted; for instance, the lab, clinic, or public areas. “Study Type” identifies research method used to validate the display. “Results Reported” summarizes key findings from the study, and “Realism” presents the latency of the display as well as the ability of the display to mimic real-world dynamism. “Cognitive Workload” reports on findings indicating reduced or increased workload and “Historic Trends” identifies displays that present historical trends that are greater than 5 minutes. The variables beginning with visual encodings for temporal trajectory, duration, and frequency identify particular techniques used by the displays to represent trends, duration of events, and frequency of events. The counter-balanced variable identifies methodologies that used strategies to minimize learning effects during the experiment. Finally, the clinical scenario variable lists the displays that were evaluated using real-world clinical scenarios. These tables, along with descriptions of the results are presented in the next section.

Comprehensive Matrix of Design Properties

The goal of the comprehensive matrix of design properties is to present design decisions that were followed to develop prototypes across all 39 papers analyzed. Visual representations were found across mainly anesthesiology (n=17), critical care (n=20), and in some multi-discipline (n=2) environments. Only one display was developed as a tool for intensive nurses [42]. Multi-discipline environments consist of 2 or more specialities, such as integrated in-patient and out-patient systems. Visual displays started to become actively contributed from the early 1990s, then increasing every 10 years, 1984 (n=1), 1985-1994 (n=8), 1995-2004 (n=13), and 2005-2016 (n=17). Integrated clinical data was also found across some displays (n=16), while a greater number of displays were devoted to the display of physiological waveforms (n=24). Number of variables presented in a single screen was wide-ranging; most displays contained greater than 20 variables per screen (n=15), followed by 11-20 variables (n=12), while 7 displays contained between 0-4 variables.

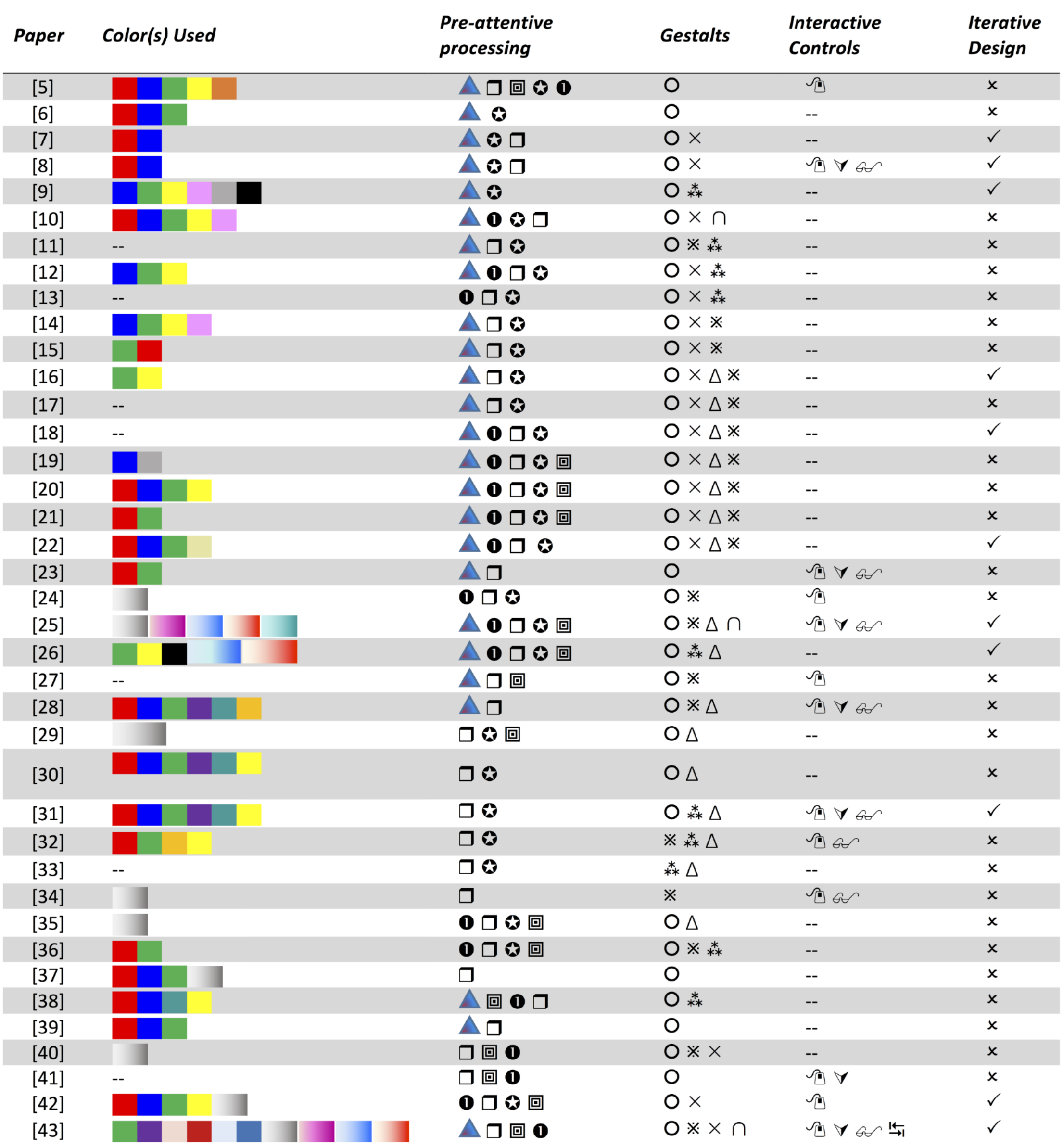
Reviewed visual representations included a mix of display formats, such as tabular (TB), object-based (OB), waveform (WF), and metaphoric (MT). Standalone MT representations were most commonly seen (n=12), followed by standalone WF (n=6). With respect to combinatory displays, TB appeared with WF (n=6) most frequently, WF with MT (n=6), and followed by WF with an OB (n=4). When identifying the display type most frequently paired in a combinatory display, WF (n=12) appeared most often, followed by TB (n=7) and OB (n=6) displays. Overall, across all identified papers including those where multiple representatives were presented, metaphoric displays were the most popular (n=22), followed by waveform displays (n=20), and object displays (n=10).

Visual representations utilized at least 2 of the primary colors, red, blue or green (n=21), while yellow (n=11) and turquoise (n=4) were also popular options. Three papers utilized discrete color encoding, 2 papers [25,43] mentioned the source of their color coding. A number of papers did not specify the type of color that was used (n=10). Pre-attentive processing of items were commonly exploited through manipulating visual variables such as color (n=24) and size (n=12), followed by dimension (n=7), and shape (n=5).

Visual representations also exploited some aspect of Gestalt’s law of groupings, such as continuity (n=18) with waveform displays, closure (n=17) when identifying boundaries, symmetry (n=14) with visual metaphors and object-based displays, and proximity (n=7) to aid in higher level detection of abnormal events. The most popular interaction method that was supported was selection (n=13). Selection allows the user to select visual objects directly to reveal greater details. This was followed by interactive filtering (n=7) to select partial ranges such as short durations of time. Finally, in many cases designs were proposed without following user-centered design approaches (n=28). In total, 10 papers reported using user-centered design processes, while 4 papers described a structured approach used in developing the proposed visual design [5,31,43,44].

XSL•FO
RenderX

Figure 1. Comprehensive Matrix of Design Properties.



Paper	*Color(s) Used*	*Pre-attentive processing*	*Gestalts*	*Interactive Controls*	*Iterative Design*
[5]	🟥🟦🟩🟨🟧	▲ ❐ ▣ ✪ ❶	O	👆	×
[6]	🟥🟦🟩	▲ ✪	O	--	×
[7]	🟥🟦	▲ ✪ ❐	O ✕	--	✓
[8]	🟥🟦	▲ ✪ ❐	O ✕	👆 ⋎ 👓	✓
[9]	🟦🟩🟨🟪⬜⬛	▲ ✪	O ⁂	--	✓
[10]	🟥🟦🟩🟨🟪	▲ ❶ ✪ ❐	O ✕ ∩	--	×
[11]	--	▲ ❐ ✪	O ※ ⁂	--	×
[12]	🟦🟩🟨	▲ ❶ ❐ ✪	O ✕ ⁂	--	×
[13]	--	❶ ❐ ✪	O ✕ ⁂	--	×
[14]	🟦🟩🟨🟪	▲ ❐ ✪	O ✕ ※	--	×
[15]	🟩🟥	▲ ❐ ✪	O ✕ ※	--	×
[16]	🟩🟨	▲ ❐ ✪	O ✕ Δ ※	--	✓
[17]	--	▲ ❐ ✪	O ✕ Δ ※	--	×
[18]	--	▲ ❶ ❐ ✪	O ✕ Δ ※	--	✓
[19]	🟦⬜	▲ ❶ ❐ ✪ ▣	O ✕ Δ ※	--	×
[20]	🟥🟦🟩🟨	▲ ❶ ❐ ✪ ▣	O ✕ Δ ※	--	×
[21]	🟥🟩	▲ ❶ ❐ ✪ ▣	O ✕ Δ ※	--	×
[22]	🟥🟦🟩🟨	▲ ❶ ❐ ✪	O ✕ Δ ※	--	✓
[23]	🟥🟩	▲ ❐	O	👆 ⋎ 👓	×
[24]	▨	❶ ❐ ✪	O ※	👆	×
[25]	▨▨▨▨▨	▲ ❶ ❐ ✪ ▣	O ※ Δ ∩	👆 ⋎ 👓	✓
[26]	🟩🟨⬛▨▨	▲ ❶ ❐ ✪ ▣	O ⁂ Δ	--	✓
[27]	--	▲ ❐ ▣	O ※	👆	×
[28]	🟥🟦🟩🟪🟦🟧	▲ ❐	O ※ Δ	👆 ⋎ 👓	×
[29]	▨	❐ ✪ ▣	O Δ	--	×
[30]	🟥🟦🟩🟪🟦🟨	❐ ✪	O Δ	--	×
[31]	🟥🟦🟩🟪🟦🟨	❐ ✪	O ⁂ Δ	👆 ⋎ 👓	✓
[32]	🟥🟩🟧🟨	❐ ✪	※ ⁂ Δ	👆 👓	×
[33]	--	❐ ✪	⁂ Δ	--	×
[34]	▨	❐	※	👆 👓	×
[35]	▨	❶ ❐ ✪ ▣	O Δ	--	×
[36]	🟥🟩	❶ ❐ ✪ ▣	O ※ ⁂	--	×
[37]	🟥🟦🟩▨	❐	O	--	×
[38]	🟥🟦🟦🟨	▲ ▣ ❶ ❐	O ⁂	--	×
[39]	🟥🟦🟩	▲ ❐	O	--	×
[40]	▨	❐ ▣ ❶	O ※ ✕	--	×
[41]	--	❐ ▣ ❶	O	👆 ⋎	×
[42]	🟥🟦🟩🟨▨	❶ ❐ ✪ ▣	O ✕	👆	✓
[43]	🟩🟪⬜🟥⬜🟦▨▨▨▨	▲ ❐ ▣ ❶	O ※ ✕ ∩	👆 ⋎ 👓 ⇆	✓

Symbols: ✓: Yes; ×: No; --: Not reported; ▲ : Color; ❐: Dimension; ▣: Size; ✪: Shape; ❶: Value; ✕: Symmetry; O: Continuity; ※: Proximity; ⁂: Similarity; Δ: Closure; ∩ : Association; 👆: Selection; ⋎: Filter; 👓: Overview; ⇆: Coordinated

XSL•FO
RenderX

Table 1. Comprehensive matrix of design properties.

Paper	Target users	Clinical context	Number of variables	Display type
Engelman et al, 2014 [6]	Intensivists	Yes	>20	Waveform display (WF)
Charabati et al, 2009 [7]	Anesthesia	No	0-4	WF
Agutter et al, 2003 [5]	Anesthesia	No	11-20	Metaphoric display (MT)
Anders et al, 2012 [8]	Intensivists	Yes	>20	WF, MT
Wachter et al, 2004 [9]	Intensivists	No	5-10	MT
van Amsterdam et al, 2013 [10]	Anesthesia	No	5-10	MT
Kennedy et al, 2011 [11]	Anesthesia	No	0-4	Object-based display (OB)
Liu et al, 2005 [12]	Intensivists	No	5-10	MT
Blikeet al, 1999 [13]	Anesthesia	No	11-20	OB
Cole et al, 1994 [14]	Intensivists	No	5-10	MT
Deneault et al, 1990 [15]	Anesthesia	No	5-10	MT
Jungk et al, 2000 [16]	Anesthesia	No	>20	OB, MT
Gurushanthaiah et al, 1995 [17]	Anesthesia	No	5-10	WF, MT
Ireland et al, 1997 [18]	Intensivists	Yes	>20	MT
Tappan et al, 2009 [19]	Anesthesia	No	5-10	MT
Michels et al, 1997 [20]	Anesthesia	No	>20	MT
Effken et al, 1997 [21]	Intensivists	No	5-10	OB, MT
Görges et al, 2012 [22]	Intensivists	Yes	11-20	WF, MT
Stylianides et al, 2011 [23]	Intensivists	Yes	>20	WF
Litt et al, 1992 [24]	Intensivists	Yes	>20	WF, MT
Gschwandtner et al, 2011 [25]	Intensivists	Yes	>20	WF, MT
Horn et al, 2001 [26]	Intensivists	Yes	11-20	MT
Dayhoff et al, 1994 [27]	Intensivists	Yes	>20	WF
Norris et al, 2002 [28]	Intensivists	Yes	>20	Tabular display (TB), WF
Langner, 1952 [29]	Intensivists	No	0-4	WF
Burykin et al, 2011 [30]	Intensivists	Yes	0-4	WF
Miller et al, 2009 [31]	Intensivists	Yes	>20	TB, WF, OB
Kruger et al, 2011 [32]	Anesthesia	Yes	11-20	MT
Law et al, 2004 [33]	Intensivists	No	5-10	TB, WF
Ahmed et al, 2011 [34]	Intensivists	Yes	>20	TB
Sainsbury, 1993 [35]	Anesthesia	No	11-20	WF, OB
Zhang et al, 2002 [36]	Anesthesia	No	5-10	OB, MT
Kennedy et al, 2008 [37]	Anesthesia	No	0-4	WF
Lowe et al, 2001 [38]	Anesthesia	No	0-4	OB
Charbonnier, 2004 [39]	Intensivists	No	0-4	TB, WF
Shabot et al, 1986 [40]	Anesthesia	No	>20	TB, MT
Eden et al, 2006 [41]	Anesthesia	Yes	>20	TB, WF, OB
Koch et al, 2013 [42]	Nurses	Yes	>20	TB, WF, MT
Kamaleswaran et al, 2016 [43]	Intensivists	Yes	11-20	WF, OB, MT

XSL•FO
RenderX

Table 2. Comprehensive matrix of study results.

Paper	Setting	Study type	Results reported	Realism	Cognitive workload	Historic trends	Visual encoding for temporal trajectory	Visual encoding for duration	Visual encoding for frequency
Engelman et al, 2014 [6]	ICU	Eval.[a]	Pos.[b]	Live	–	C[c]	C	–	–
Charabati et al, 2009 [7]	Lab	Exp.[d]	Pos.	Static	↓	–	C	–	–
Agutter et al, 2003 [5]	Lab	Exp.	Pos.	Sim.[e]	↓	–	–	–	–
Anders et al, 2012 [8]	ICU	Exp.	±	Static	0[f]	C	C	–	–
Wachter et al, 2004 [9]	ICU	Eval.	Pos.	Live	–	–	G[g]	–	–
van Amsterdam et al, 2013 [10]	Lab	Exp.	Neg.[h]	Static	–	–	O[i]	–	–
Kennedy et al, 2011 [11]	Lab	Exp.	Pos.	Sim.	–	–	O	–	–
Liu et al, 2005 [12]	Lab	Exp.	Pos.	Sim.	–	–	–	–	–
Blikeet al, 1999 [13]	Lab	Exp.	Pos.	Sim.	–	–	–	–	–
Cole et al, 1994 [14]	Lab	Exp.	Pos.	Static	–	–	G	G	G
Deneault et al, 1990 [15]	Lab	Exp.	Pos.	Sim.	–	–	–	–	–
Jungk et al, 2000 [16]	Lab	Exp.	Pos.	Sim.	–	–	C	–	–
Gurushanthaiah et al, 1995 [17]	Lab	Exp.	Pos.	Sim.	–	–	C	–	–
Ireland et al, 1997 [18]	Lab	Eval.	Pos.	Static	–	–	C, G	–	–
Tappan et al, 2009 [19]	Lab	Exp.	Pos.	Sim.	↓	–	C, G	–	–
Michels et al, 1997 [20]	Lab	Exp.	Pos.	Sim.	–	–	–	–	–
Effken et al, 1997 [21]	Lab	Exp.	Pos.	Sim.	–	–	–	–	–
Görges et al, 2012 [22]	ICU	Exp.	Pos.	Sim.	0	–	G	G	G
Stylianides et al, 2011 [23]	ICU	Eval.	Pos.	Live	–	C	C	–	–
Litt et al, 1992 [24]	Lab	App.[j]	–	Static	–	C	C	G	G
Gschwandtner et al, 2011 [25]	Lab	Des.[k]	–	Static	–	C	C	C	–
Horn et al, 2001 [26]	ICU	Eval.	Pos.	Static	↓	C	G	G	–
Dayhoff et al, 1994 [27]	ICU	App.	–	Live	–	–	C	–	–

XSL•FO
RenderX

Paper	Setting	Study type	Results reported	Realism	Cognitive workload	Historic trends	Visual encoding for temporal trajectory	Visual encoding for duration	Visual encoding for frequency
Norris et al, 2002 [28]	ICU	App.	Pos.	Live	–	C	C	–	–
Langner, 1952 [29]	ICU	Eval.	–	Static	–	–	C	–	–
Burykin et al, 2011 [30]	ICU	App.	–	Sim.	–	–	C	–	–
Miller et al, 2009 [31]	ICU	Exp.	Pos.	Static	–	–	C	T[l]	T
Kruger et al, 2011 [32]	Surgery	App.	–	Live	–	–	G	T	T
Law et al, 2004 [33]	Lab	Exp.	–	Static	–	–	C, T	T	T
Ahmed et al, 2011 [34]	Lab	Exp.	Pos.	Sim.	↓	–	T	T	T
Sainsbury, 1993 [35]	Surgery	Eval.	Pos.	Live	–	–	C	–	–
Zhang et al, 2002 [36]	Lab	Exp.	±	Sim.	±	–	C, G	G	G
Kennedy et al, 2008 [37]	Lab	Exp.	Pos.	Sim.	–	–	C	–	–
Lowe et al, 2001 [38]	Lab	App.	Pos.	Sim.	–	–	C	–	–
Charbonnier, 2004 [39]	Lab	Des.	–	Sim.	–	–	C	–	–
Shabot et al, 1986 [40]	Lab	Des.	–	Sim.	–	C	C	–	–
Eden et al, 2006 [41]	Surgery	App.	Pos.	Live	↓	C	C	–	–
Koch et al, 2013 [42]	ICU	Exp.	Pos.	Sim.	↓	C	C	–	–
Kamaleswaran et al, 2016 [43]	ICU	Eval.	Pos.	Live	–	C, G	C, G, O	G, O	G, O

[a] Eval: Evaluation

[b] Pos: Positive

[c] C: Curves

[d] Exp: Experiment

[e] Sim: Simulated

[f] 0: No Change

[g] G: Glyph

[h] Neg: Negative

[i] O: Object

[j] App: Application

[k] Des: Design

[l] T: Text

XSL•FO
RenderX

Table 3. Comprehensive matrix of study results.

Paper	Counter-balanced for display or scenario	Scenarios were clinically relevant	Function supporting case-controlled analysis
Engelman et al, 2014 [6]	–	Yes	–
Charabati et al, 2009 [7]	Display	Yes	–
Agutter et al, 2003 [5]	Display	Yes	–
Anders et al, 2012 [8]	Display & Scenario	Yes	–
Wachter et al, 2004 [9]	–	No	–
van Amsterdam et al, 2013 [10]	Display	Yes	–
Kennedy et al, 2011 [11]	Display	No	–
Liu et al, 2005 [12]	Display & Scenario	Yes	–
Blikeet al, 1999 [13]	Scenario	Yes	–
Cole et al, 1994 [14]	Display & Scenario	Yes	–
Deneault et al, 1990 [15]	Display & Scenario	Yes	–
Jungk et al, 2000 [16]	Scenario	Yes	–
Gurushanthaiah et al, 1995 [17]	Scenario	Yes	–
Ireland et al, 1997 [18]	–	No	–
Tappan et al, 2009 [19]	Display & Scenario	Yes	–
Michels et al, 1997 [20]	Display & Scenario	Yes	–
Effken et al, 1997 [21]	Scenario	Yes	–
Görges et al, 2012 [22]	Scenario	Yes	–
Stylianides et al, 2011 [23]	–	No	–
Litt et al, 1992 [24]	–	No	–
Gschwandtner et al, 2011 [25]	–	Yes	✓
Horn et al, 2001 [26]	–	No	–
Dayhoff et al, 1994 [27]	–	No	–
Norris et al, 2002 [28]	–	No	–
Langner, 1952 [29]	–	No	–
Burykin et al, 2011 [30]	–	No	–
Miller et al, 2009 [31]	Display & Scenario	Yes	–
Kruger et al, 2011 [32]	–	No	–
Law et al, 2004 [33]	Display & Scenario	Yes	–
Ahmed et al, 2011 [34]	Display & Scenario	Yes	–
Sainsbury, 1993 [35]	–	No	–
Zhang et al, 2002 [36]	Scenario	Yes	–
Kennedy et al, 2008 [37]	Display & Scenario	No	–
Lowe et al, 2001 [38]	–	No	–
Charbonnier, 2004 [39]	–	No	–
Shabot et al, 1986 [40]	–	No	–
Eden et al, 2006 [41]	–	No	–
Koch et al, 2013 [42]	Display & Scenario	Yes	–
Kamaleswaran et al, 2016 [43]	Display & Scenario	Yes	–

XSL•FO
RenderX

Comprehensive Matrix of Study Design

The Comprehensive Matrix of Study Design (Figure 2) presents the results that were reported by authors for any evaluation or experiment. While the search strategy yielded 39 full papers that were identified for analysis, only 29 of these papers contained primary study results from a case study, evaluation, or human experiment, and employed 1 of naïve, novice, or expert participants in the evaluation method. Naïve participants were generally undergraduate students with little or no prior clinical knowledge. Novice participants ranged from undergraduate computer science or nursing students to newly graduated clinical staff. Expert participants had at least 10 years of experience.

The number of participants exposed to test conditions highly varied; however, the majority of studies employed at least 15 participants. Six studies used a sample size greater than 20 to test for detection, diagnostic, and treatment accuracy, with the minimum and maximum being 4 and 32 participants, respectively. Most displays integrated these systems in a single display using live or static representations (n=15), while displays that were presented as case studies (in situ) were connected to central monitoring systems. Some displays supported views of clinical information that integrated data from other clinical and laboratory systems (n=15) [45]. Most prototypes that were evaluated used more than one data stream, with the exception of the studies that contained low-frequency updates (n=9). Most evaluations or experiments utilized more than one condition to test each display; however, a few did not have any scenarios or patient conditions (n=9). A large number of studies also did not utilize data from more than one patient-source (n=26).

Most of the studies were conducted in laboratory environments (n=24), followed by evaluations or experiments in the intensive care unit (n=12). Some studies were evaluated over multiple specialities (n=2). A majority of studies used some form of experimentation to validate their designs (n=21), although the specific method of experimentation was not always explicitly mentioned. Evaluations involved clinicians and mixed qualitative and quantitative methods were used to report results (n=8). Applications were primarily qualitative in nature, often depicting results through anecdotes (n=7). The remaining studies were design papers that investigated novel visual representations without involving prototypes. Of the papers that reported results (n=31), most reported positive findings (n=27), but in some cases negative results were also reported (n=4). A between-group experimental study yielded site-dependent results that were skewed towards the site that produced the visual representation. For evaluations or experiments the source of data to support realism was spread across live simulations (n=19), live patient-origin data (n=9), or static patient-generated data (n=11). Most studies did not test for cognitive overload using ad hoc methods or traditional workload score metrics such as the NASA Task Load Index (NASA-TLX) (n=30). Where cognitive workload was reported (n=7), most were reported to have reduced cognitive overload (n=5), while others reported no change or mixed results (n=3).

Long-term historical values, specifically ranges exceeding 5 minutes of monitoring were not included in majority of the displays (n=28). Tri-event parameters, namely, trajectory, frequency, and duration were seldom supported by visual representations, where these parameters were identified, trajectory was most frequently found (n=27). Temporal trajectory was encoded using curves (n=25) such as in a line graph, or glyphs (n=9). In terms of duration, the second tri-event parameter was seen across 10 displays, of which, glyphs (n=6), text (n=4) or curves (n=2) representations were utilized. Frequency, the last tri-event parameter was also seen in some visual representations encoded by glyph (n=5) or text (n=4) where supported. Where displays were validated through experimentation, both the display and scenarios were more often counterbalanced (n=12), while some experiments counterbalanced only the scenario (n=6) and others only the display (n=4). Scenarios were utilized across many studies utilizing experimentation or evaluation methodologies (n=22) and most were clinically relevant problems (n=21). Finally, only one of the evaluated visual representations supported the ability to perform analysis across multiple patients.

Discussion

Principal Findings

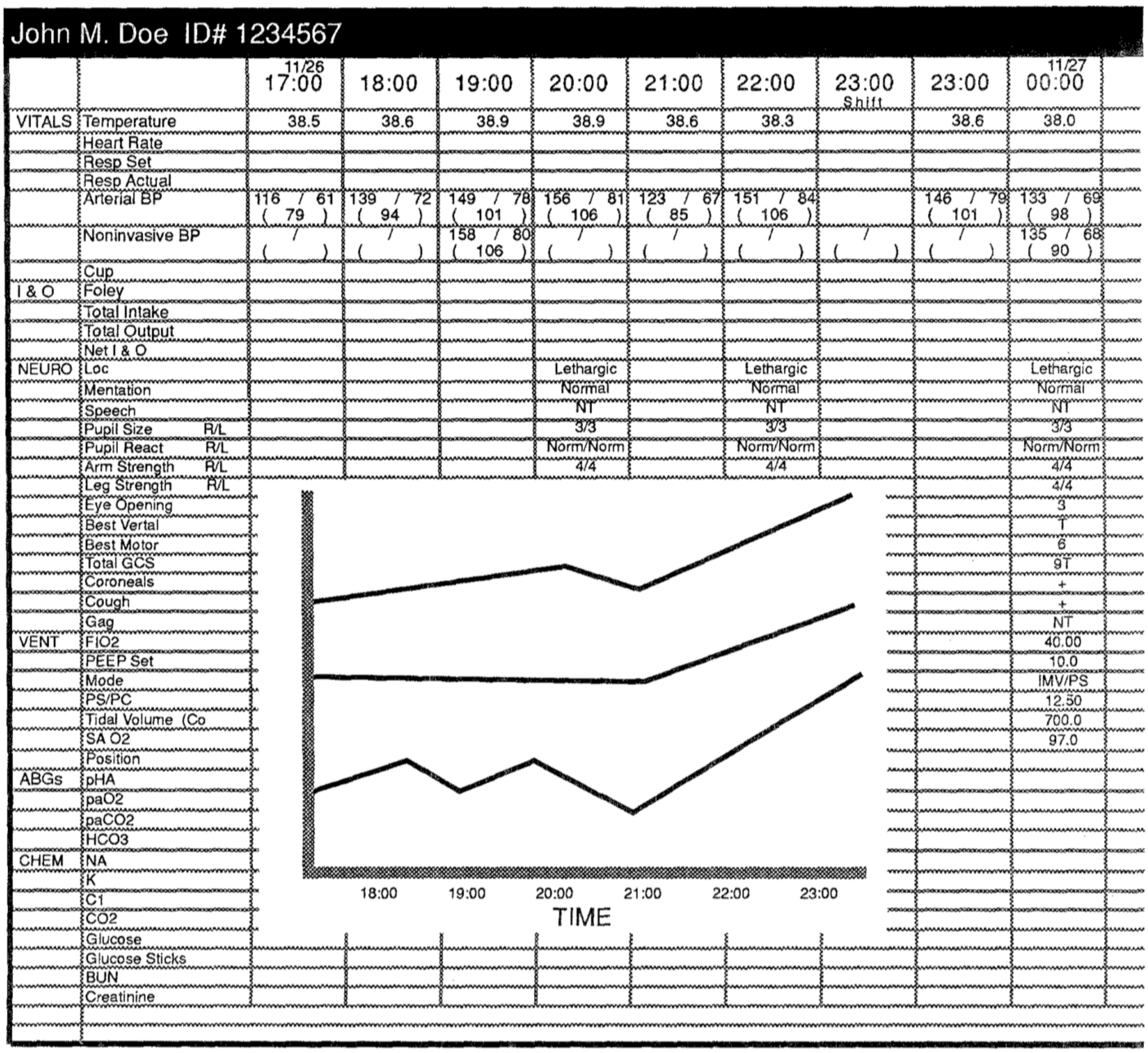
A total of 19 novel visual representations were identified from analysis of the literature. Novel displays were seen across 4 main groups, including tabular, waveform (graph-based), object, and metaphors. The latter 2 are aggregated together as ecologic displays.

Tabular Displays

The early- 1990's saw growing interest in converting large volumes of paper patient charts to "virtual" records [24,46-49]. Initial representations adopted by these virtual patient records were largely tabular and text-dominant, and sometimes contributed negatively to information overload [48]. Figure 2 [50] illustrates an example of a traditional virtual patient chart that mimics a traditional paper flow chart. This review identified 14 tabular representations published from 1952 to 1997. Those systems provide a direct manipulation using the traditional desktop-oriented, Windows-Icon-Mouse-Pointer (WIMP) interaction paradigm. Additional levels of interactions, such as multiple mouse clicks, are required to access unique views of patient data. Large number of these displays are often duplicated to a physical copy, in part due to the simplicity and ease of reading paper charts [51].

XSL•FO
RenderX

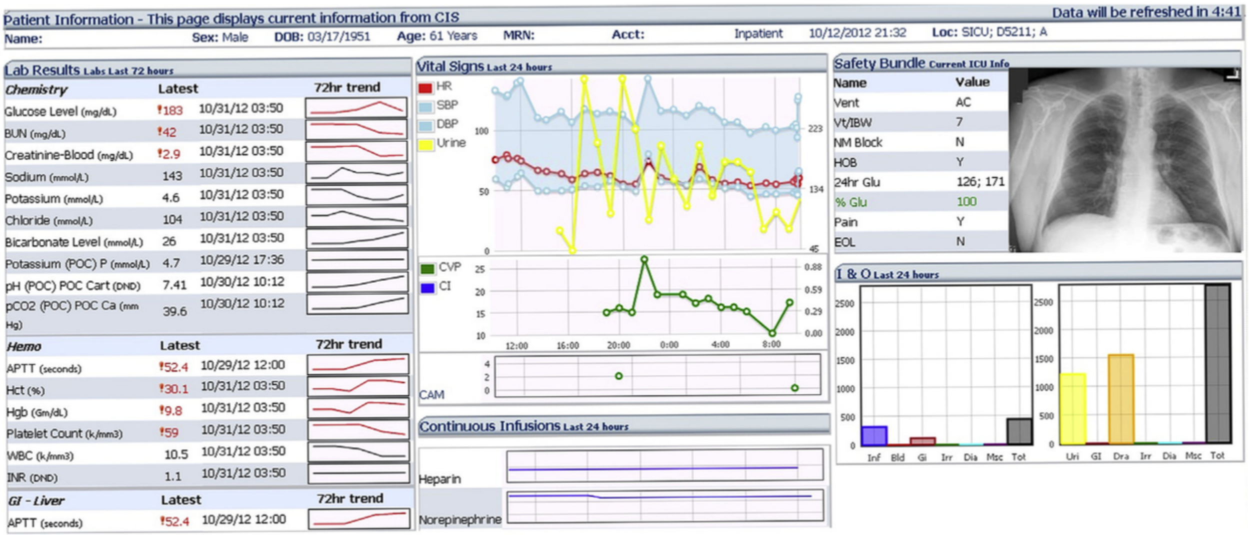
Figure 2. A tabular display that mimics traditional clinical flow-sheets.



John M. Doe ID# 1234567

		11/26 17:00	18:00	19:00	20:00	21:00	22:00	23:00 Shift	23:00	11/27 00:00
VITALS	Temperature	38.5	38.6	38.9	38.9	38.6	38.3		38.6	38.0
	Heart Rate									
	Resp Set									
	Resp Actual									
	Arterial BP	116 / 61 (79)	139 / 72 (94)	149 / 78 (101)	156 / 81 (106)	123 / 67 (85)	151 / 84 (106)		146 / 79 (101)	133 / 69 (98)
	Noninvasive BP	/ ( )	/ ( )	158 / 80 (106)	/ ( )	/ ( )	/ ( )	/ ( )	/ ( )	135 / 68 (90)
	Cup									
I & O	Foley									
	Total Intake									
	Total Output									
	Net I & O									
NEURO	Loc				Lethargic		Lethargic			Lethargic
	Mentation				Normal		Normal			Normal
	Speech				NT		NT			NT
	Pupil Size R/L				3/3		3/3			3/3
	Pupil React R/L				Norm/Norm		Norm/Norm			Norm/Norm
	Arm Strength R/L				4/4		4/4			4/4
	Leg Strength R/L									4/4
	Eye Opening									3
	Best Vertal									T
	Best Motor									6
	Total GCS									9T
	Coroneals									+
	Cough									+
	Gag									NT
VENT	FiO2									40.00
	PEEP Set									10.0
	Mode									IMV/PS
	PS/PC									12.50
	Tidal Volume (Co									700.0
	SA O2									97.0
	Position									
ABGs	pHA									
	paO2									
	paCO2									
	HCO3									
CHEM	NA									
	K									
	C1									
	CO2									
	Glucose									
	Glucose Sticks									
	BUN									
	Creatinine									

Figure 3. A modern dashboard utilizing waveform displays.



Patient Information - This page displays current information from CIS

Data will be refreshed in 4:41

Name: **Sex:** Male **DOB:** 03/17/1951 **Age:** 61 Years **MRN:** **Acct:** Inpatient 10/12/2012 21:32 **Loc:** SICU; D5211; A

Lab Results Labs Last 72 hours

Chemistry	Latest		72hr trend
Glucose Level (mg/dL)	!183	10/31/12 03:50	
BUN (mg/dL)	!42	10/31/12 03:50	
Creatinine-Blood (mg/dL)	!2.9	10/31/12 03:50	
Sodium (mmol/L)	143	10/31/12 03:50	
Potassium (mmol/L)	4.6	10/31/12 03:50	
Chloride (mmol/L)	104	10/31/12 03:50	
Bicarbonate Level (mmol/L)	26	10/31/12 03:50	
Potassium (POC) P (mmol/L)	4.7	10/29/12 17:36	
pH (POC) POC Cart (DND)	7.41	10/30/12 10:12	
pCO2 (POC) POC Ca (mm Hg)	39.6	10/30/12 10:12	
Hemo	**Latest**		**72hr trend**
APTT (seconds)	!52.4	10/29/12 12:00	
Hct (%)	!30.1	10/31/12 03:50	
Hgb (Gm/dL)	!9.8	10/31/12 03:50	
Platelet Count (k/mm3)	!59	10/31/12 03:50	
WBC (k/mm3)	10.5	10/31/12 03:50	
INR (DND)	1.1	10/31/12 03:50	
GI - Liver	**Latest**		**72hr trend**
APTT (seconds)	!52.4	10/29/12 12:00	

Vital Signs Last 24 hours

Safety Bundle Current ICU Info

Name	Value
Vent	AC
Vt/IBW	7
NM Block	N
HOB	Y
24hr Glu	126; 171
% Glu	100
Pain	Y
EOL	N

I & O Last 24 hours

Continuous Infusions Last 24 hours

XSL•FO
RenderX

Waveform Displays

The review identified 9 out of 39 studies that used some form of live physiologic streams from real patients to display largely identical waveform representations. It was also noted that much of these waveform displays were integrated with other tabular and text representations. Five papers that presented waveform displays also supported interactive capabilities, including the ability to select regions of interest, filter based on patients, and generate screen captures [6,8,25,43,52]. Stylianides and colleagues (2011) present an engine for producing waveform graphics [23]; however, their system serves the purpose of animating historic physiologic data streams. CareCruiser [25], supports the interactive exploration of treatment plans using physiologic data. However, that system was not evaluated using more than one clinical user. PhysioEx [43] was evaluated using an expert evaluation methodology employing 5 domain experts, and was shown to further enhance that interactive analytic workflow by providing coordinated analysis of temporal data streams; however, using waveform displays only to guide the user with additional context.

Despite their ability to communicate acute time-sensitive events [20], waveform representations have numerous limitations [4,31,53]. One prime disadvantage of waveform displays is the potential to negatively impact cognitive load, that is, they require humans to monitor and consume large numbers of data points as they are produced to derive trends and higher level knowledge [7,8,22]. These waveforms display can convey several features in one frame; therefore, easily disturb limited resources of the working memory capabilities [54]. The challenge of managing large volumes of data have been extensively studied in several domains, such as information overload [55], visual data mining [56], and addressing cognitive challenges related to interruptions, task performance, and decision making [55,57-59].

Integrated methods of representing critical physiological information have been actively studied to reduce the internal mental processing requirement [20,22,32,60-62]. These integrated displays use a combination of text [33,34], graphic [3,4,63], and waveform [64,65] representations to summarize low-level information. Figure 3 [6] illustrates an example of an integrated display. Three such integrated displays were identified in the review [6,8,25]. These displays support clinicians to interactively select regions of interest while monitoring other forms of slow-changing clinical data. However, only one display allows the clinician to compare against a cohort [25]. Other studies, seeking alternatives to the waveform visual encoding, propose novel and ecological methods to improve knowledge discovery and minimize cognitive overload.

Ecological Displays

Classes of Visual Representations

Ecologic displays attempt to integrate relationships existing across both workflows and semantics [66]. Among the primary goals of ecologic displays is to convey both the means-end relation, answering the particular means of arriving at that state and its ultimate consequence. From our review, we identified 2 large classes of visual representations that approach these objectives. Object-oriented displays, and metaphoric displays were seen to extend typical limitations found in text, tabular, and waveform displays by introducing novel information, such as spatial and temporal arrangements of closely related information.

Object-Oriented Displays

Displays that utilize and manipulate 2-dimensional graphical objects, limited to basic shapes and symmetries to produce emergent properties have been classed as object-oriented displays [2,13,35]. These displays follow demonstrated efficacy of graphical displays over traditional numeric displays observed in nuclear power station control stations [67]. Studies have shown a positive relationship with integrated displays and an overall improvement in diagnosis ability as well as a reduction in time to initiate treatment [68].

Blike and colleagues (2000) [69] showed that subjects exposed to emergent features using novel graphics recognized a problem more rapidly, but their accuracy had not improved in comparison to the numeric display. Moreover, they showed that the shape of the graphic, illustrated in Figure 4 [13], improved detection of etiology compared to the numeric and control displays. While Blike and colleagues stated an improved reaction and fewer errors when using the object-oriented display, the display was found to be confusing and not ecological to naïve participants. Zhang and colleagues [36] reproduced the designs introduced by Blike et al, and found that anesthesiologists were able to detect simple deviations faster; however, no change was seen with detection times of more complex cardiovascular events. Other studies have reported similar conclusions [5,9,13,68,70], suggesting a link between detection and reactionary time to the format and features of the graphical display.

In contrast, other studies that extrapolated heuristics from object-oriented displays report less convincing evidence; for instance, some report negative links when participants were presented object-oriented displays [21,37]. The etiological potential display (Figure 5) [21] attempts to extract specific features of object displays that improve detection and diagnosis. In that study, Effken and colleagues find no significance in the detection or diagnostic times, even when 3 abstract displays were tested. Two of these displays required that features of the full prototype either be reorganized or removed.

XSL•FO
RenderX

Figure 4. Advanced graphical display for hemodynamic monitoring.

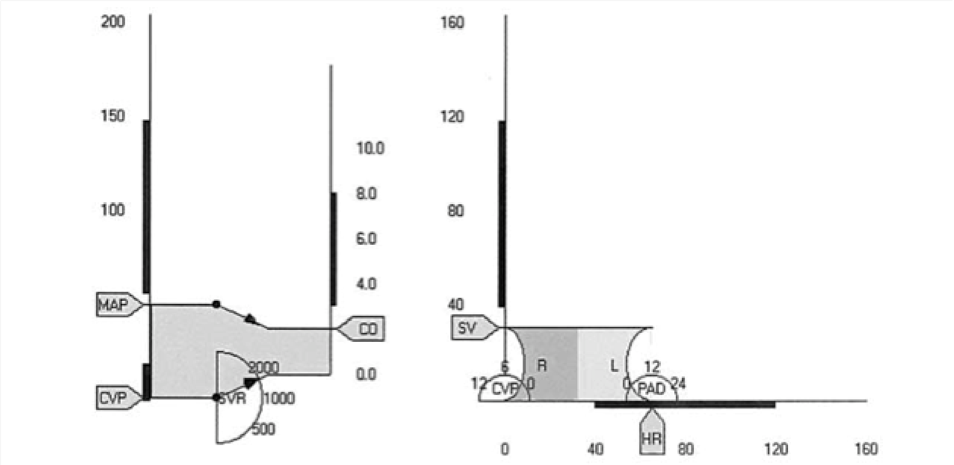
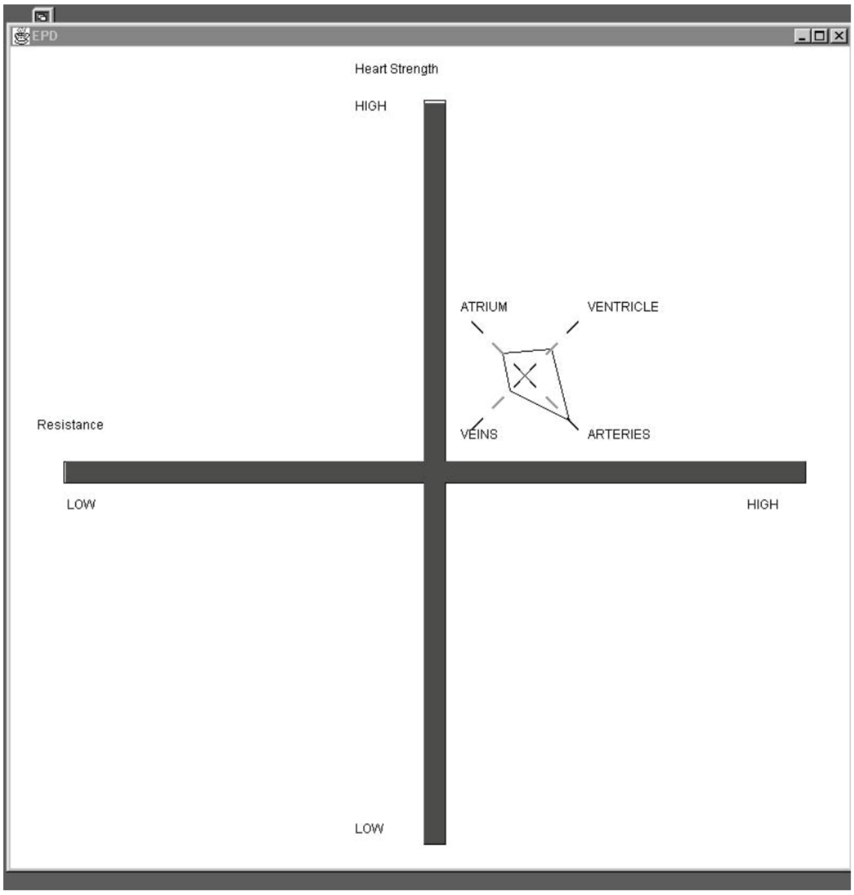


Figure 5. The etiological potential display moves an object across 4 quadrants of heart strength and resistance. The object in the top right quadrant is distorted to show relative depressions in the atrium, ventricles, veins and arteries.



Metaphoric Displays

A total of 20 representations, for over half of all visual representations analyzed, belonged to the metaphoric display group. Most clinical metaphoric displays illustrate physiologic data in terms of organ-systems [20,36,44,71]. Five papers presented metaphors that involved dynamic objects that exhibited behaviors similar to organ systems [5,14,20,21,26].

Several papers identified metaphor displays with positive outcomes. Cole and Stewart (1994) [14], introduced a visual representation (Figure 6) [50] which consists of 2 volume rectangles that compress or expand similar to the respiratory system. This design was further improved with additional data dimensions [26]. A Graphical Cardiovascular Display (Figure 7) [5] that uses a pipe-like metaphor of the cardiovascular system, was shown to enable faster detection of adverse events [5]. Wachter and colleagues (2003) applied similar approaches to develop a respiratory interface and found participants were able to identify abnormal states faster [9]. Gorges and colleagues, introduced a series of visual metaphors to communicate visual signs to bed-side clinicians [22]. These displays adopt a clock metaphor illustrated in Figure 8 [22] to

XSL•FO
RenderX

convey salient features, such as temporal trends over the past 12-hours. Charabati and colleagues from the Montreal General Hospital's department of anesthesiology introduced a gauge metaphor to highlight normal and abnormal ranges, and conducted an evaluation across 2 sites [7]. They found a combination of numeric and visual metaphors achieved the strongest advantage in detection, accuracy, and workload. Tappan and colleagues evaluated visual metaphors by appending visual objects to traditional medical monitors [19]. They reported significant improvements in detection of adverse events, with the visual metaphor having a 14.4 second advantage over traditional physiologic monitors. The visual metaphor was also found to reduce the number of missed events. However, similar to previous studies, these investigations were conducted in controlled environments.

Not all visual metaphors, however, have seen similar success. Zhang and colleagues (2002) [36] introduced an integrated 3-dimensional balloon metaphor, building on the work of Blike and colleagues (2000) [69] with object displays. Zhang and colleagues found mixed results after evaluations, with only 63% of scenarios having shorter detection than scenarios, and situational awareness being improved in 1 of 4 scenarios. Moreover, van Amsterdam and colleagues (2013) from the University Medical Center Groningen, utilized customization features offered by vendor-based medical monitors to construct and evaluate a metaphoric display presented in Figure 9 [10]. They found, however, that visual metaphors did not improve detection or accuracy of anesthesiologists [10].

Finally, while ecologic representations were evaluated for diagnostic accuracy and speed, the challenges surrounding cognitive errors remain only a secondary concern in research involving visual representations. Less than 8 out of 39 of papers analyzed were identified to have measured for cognitive workload [5,7,8,19,22,34,36,41]. Of the 8 papers that measured for cognitive workload, 4 papers used a quantitative measure such as the NASA-TLX score [5,8,22,34]. There are also limitations with the use of NASA-TLX, largely because it is a self-reported method of identifying perceived workload. A total of 3 of the 8 papers were evaluated with critical care clinicians, consequently, incorporating cognitive workload as a passive measure of potential cognitive error remains limited across visual representation research for clinical environments. Significantly, none of metaphoric displays supported analytic functions.

Figure 6. Volume triangles represent multivariate clinical data using a lung-expansion metaphor.

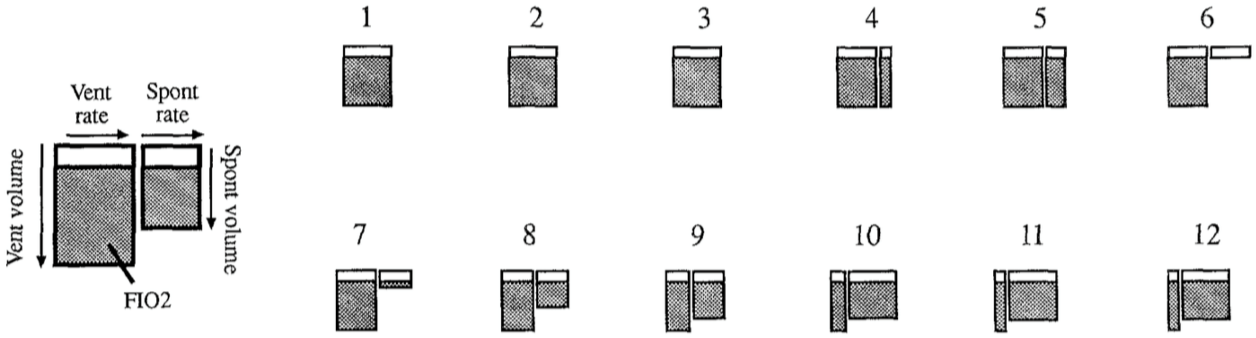
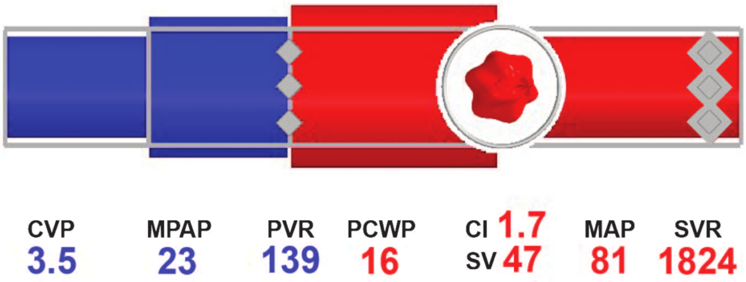


Figure 7. Graphical Cardiovascular Display, adapts a metaphor of a pipes with volume and pressure properties.



XSL•FO
RenderX

Figure 8. Far-view visual metaphors for triaging vital signs.

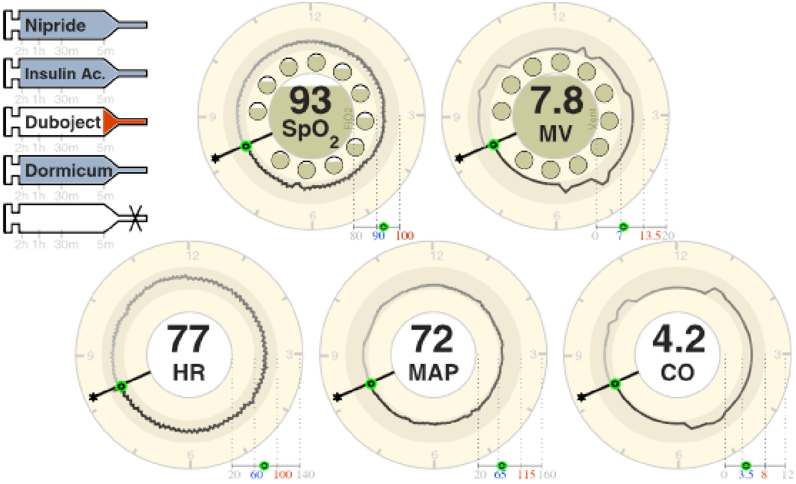
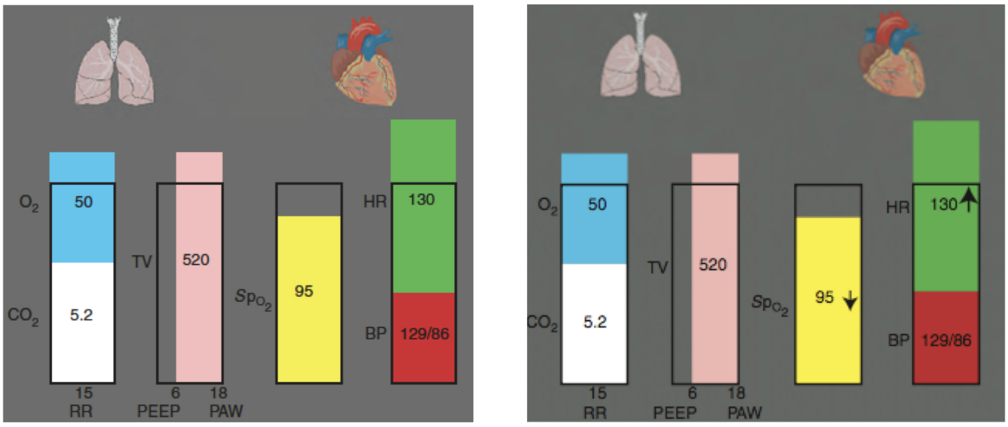


Figure 9. (a) Metaphorical anesthesia interface and (b) Metaphorical interface with trend information (tMAI).



(a) (b)

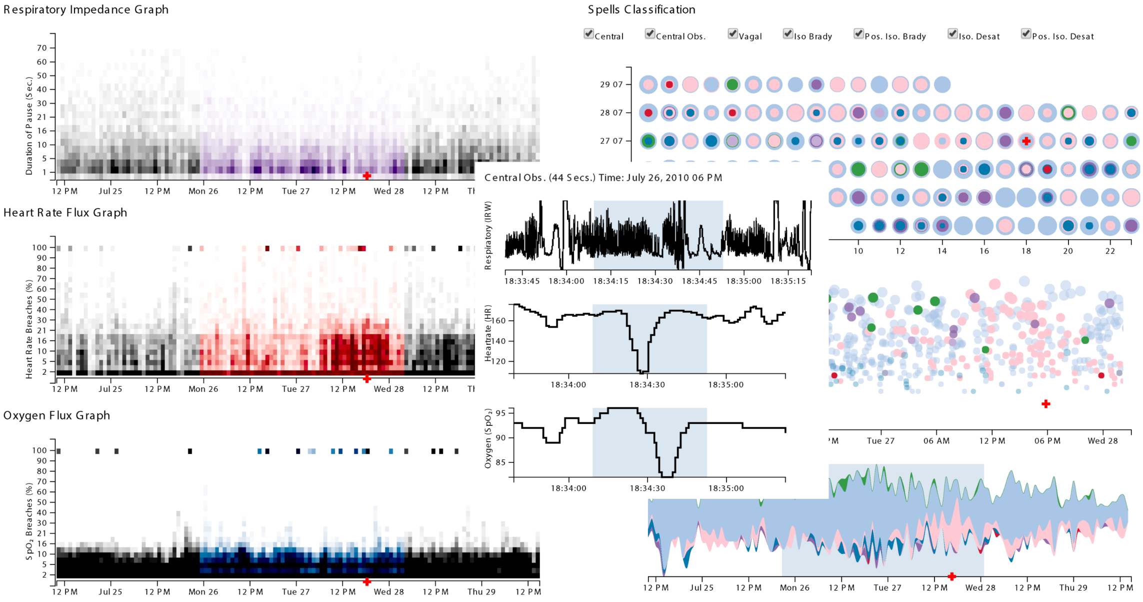
Tri-Event Parameters

Physiological displays can be designed and developed using 3 consumption efficacy metrics derived for temporal and dynamic data streams [43]. These metrics are termed tri-event temporal parameters namely, trajectory, frequency, and duration of salient events.

Among the tri-event parameters, trajectory was found to be the most popular, with 32 of 39 studies incorporating some form of trajectory information. However, longitudinal trajectory was found in only 9 studies, and was rare among displays that were found in anesthesiology but more common in critical care. Displays that incorporated an aspect of the tri-event temporal parameters exclusively adopted trajectory. Nine visual representations were found to have included the duration and frequency metrics. Most of the representations that included duration and frequency used glyphs (n=6) or text (n=5) to communicate episodic information. For instance, PhysioEx [43] uses the river metaphor [72] to illustrate frequency of adverse physiologic events that were analyzed by a real-time algorithm (bottom left, in Figure 10). Text also remains a popular method for communicating discrete events. Law and colleagues found text to be superior to waveform and numeric displays when communicating clinical episodes, even while clinicians reported a preference for graphical displays [73]. Where multiple views were presented, only one representation utilized interactive coordination between independent views [6].

XSL•FO
RenderX

Figure 10. PhysioEx, a coordinated visual analytic tool for exploring clinical events across multiple temporal physiologic data streams.



Conclusion

Visual representations of physiological data have been attempted several times as witnessed by the sheer size of prior work discussed in this paper. Many have shown their potential to improve clinical care, and while largely positive results have been released, there are still concerns as to the efficacy of both in reproducibility as well as translatability to the unit. In particular, methods to identify the accuracy of actions post-treatment to the display remain as concern and open areas for further exploration.

Few clinical visualization papers studied associations of the treatment condition to the accuracy or accrued insight by the user. It was also seen that most studies included detailed study of the time to diagnosis and its accuracy; however, many of these studies included highly controlled scenarios with highly visible graphical distortions. Additionally, few studies used real patient data to evaluate their prototypes. Hence, the frequency of events with clear and distinctive graphical patterns existing across real patient data remains untested. Detection was also another area where studies frequently report positive findings; however, in many cases these differences were marginal and found in narrow statistical ranges. It has yet to be proven whether these statistical significances are relevant in the clinical domain. Exact mechanisms inducing positive effect have yet to be studied within the prototypes studied [63,74].

Visual representations show promise; however, they are plagued with user-preference and interaction challenges. Results spanning two decades continue to show positive influence of graphical representations when they are used in simulated studies [4]. However, many of these studies have not used standardized metrics to test distinct controlled variables, or provide evidence of precisely which features of the graphical displays afford greater comprehension to the consumer. Questions still remain as to its efficacy in clinical practice, where, the availability of all data required by the representations may be limited. There is also the limitation of graphical representation failing to maintain interpretable coherence, when provided incorrect data [2].

Some studies have also demonstrated user involvement as an important factor which may have influenced results, in the design and development of the clinical system [45]. Future studies should focus on clinical validation as a means to identify real-life relevance. Clinical experiments are difficult in lieu of several considerations and their limitations. However, one study by Wachter et al [9], demonstrates that observational studies, although somewhat intrusive, may produce some significant qualitative results. These studies need to be expanded, and clinical trials must ultimately demonstrate their efficacy. Cognitive errors also require additional research effort, specifically by including evaluation methodologies such as the NASA-TLX score to allow end-users to self-report perceived workloads.

Only 7 visualizations were identified to have had some element of interactive selection and filtering functions to support basic analysis tasks. While only one display was identified to support analysis across cohort populations. The general absence of analysis functionalities is an opportunity for enhancing physiologic visualizations. Physiologic data represents a unique subset, due to the dynamic and streaming nature of the data. Application of visual analysis techniques may support novel uses of physiologic visualizations, such as supporting human-driven hypothesis generation tasks.

Finally, research in visual representations should include tri-event parameters as important design considerations to produce designs that communicate episodic information. PhysioEx was seen to incorporate all 3 parameters; however, it was limited to one view per patient [43]. These visual representations can then be used to better assess the influence of tri-event parameters on higher level workflows as well as in the progression of clinical conditions.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. McGregor C. Big data in neonatal intensive care. Computer 2013 Jun;46(6):54-59. [doi: 10.1109/MC.2013.157]
2. Sanderson PM, Watson MO, Russell WJ. Advanced patient monitoring displays: tools for continuous informing. Anesth Analg 2005 Jul;101(1):161. [doi: 10.1213/01.ANE.0000154080.67496.AE] [Medline: 15976225]
3. Drews FA, Westenskow DR. The right picture is worth a thousand numbers: data displays in anesthesia. Hum Factors 2006;48(1):59-71. [Medline: 16696257]
4. Görges M, Staggers N. Evaluations of physiological monitoring displays: a systematic review. J Clin Monit Comput 2008 Feb;22(1):45-66. [doi: 10.1007/s10877-007-9106-8] [Medline: 18064532]
5. Agutter J, Drews F, Syroid N, Westneskow D, Albert R, Strayer D. Evaluation of graphic cardiovascular display in a high-fidelity simulator. Anesth Analg 2003 Nov;97(5):1403-1413 [FREE Full text]
6. Engelman D, Higgins T, Talati R, Grimsman J. Maintaining situational awareness in a cardiac intensive care unit. J Thorac Cardiovasc Surg 2014 Mar;147(3):1105-1106 [FREE Full text] [doi: 10.1016/j.jtcvs.2013.10.044]
7. Charabati S, Bracco D, Mathieu P, Hemmerling T. Comparison of four different display designs of a novel anaesthetic monitoring system, the 'integrated monitor of anaesthesia (IMA)'. Br J Anaesth 2009 Nov;103(5):670-677 [FREE Full text] [doi: 10.1093/bja/aep258] [Medline: 19767312]
8. Anders S, Albert R, Miller A, Weinger M, Doig A, Behrens M, et al. Evaluation of an integrated graphical display to promote acute change detection in ICU patients. Int J Med Inform 2012 Dec;81(12):842-851 [FREE Full text] [doi: 10.1016/j.ijmedinf.2012.04.004] [Medline: 22534099]
9. Wachter S, Markewitz B, Rose R, Westenskow D. Evaluation of a pulmonary graphical display in the medical intensive care unit: an observational study. J Biomed Inform 2005 Jun;38(3):239-243 [FREE Full text] [doi: 10.1016/j.jbi.2004.11.003] [Medline: 15896697]
10. van Amsterdam K, Cnossen F, Ballast A, Struys M. Visual metaphors on anaesthesia monitors do not improve anaesthetists' performance in the operating theatre. Br J Anaesth 2013 Feb 05;110(5):816-822 [FREE Full text] [doi: 10.1093/bja/aes516] [Medline: 23384736]
11. Kennedy R, Merry A. The effect of a graphical interpretation of a statistic trend indicator (Trigg's Tracking Variable) on the detection of simulated changes. Anaesth Intensive Care 2011 Sep;39(5):881-886 [FREE Full text] [Medline: 21970133]
12. Liu Y, Osvalder AL. Usability evaluation of a GUI prototype for a ventilator machine. J Clin Monit Comput 2004 Dec;18(5-6):365-372. [Medline: 15957628]
13. Blike GT, Surgenor SD, Whalen K. A graphical object display improves anesthesiologists' performance on a simulated diagnostic task. J Clin Monit Comput 1999 Jan;15(1):37-44. [Medline: 12578060]
14. Cole WG, Stewart JG. Human performance evaluation of a metaphor graphic display for respiratory data. Methods Inf Med 1994 Oct;33(4):390-396. [Medline: 7799815]
15. Deneault L, Lewis C, Debons A, Stein K, Dewolf A. An integrative display for patient monitoring. Presented at: IEEE International Conference on Systems, Man and Cybernetics; 1990 Nov; Los Angeles, CA. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=142161 [doi: 10.1109/ICSMC.1990.142161]
16. Jungk A, Thull B, Hoeft A, Rau G. Evaluation of two new ecological interface approaches for the anesthesia workplace. J Clin Monit Comput 2000;16(4):243-258. [Medline: 12578071]
17. Gurushanthaiah K, Weinger M, Englund C. Visual display format affects the ability of anesthesiologists to detect acute physiologic changes: a laboratory study employing a clinical display simulator. Anesthesiology 1995 Dec;83(6):1184-1193. [Medline: 8533911]
18. Ireland R, James HV, Howes M, Wilson A. Design of a summary screen for an ICU patient data management system. Med Biol Eng Comput 1997 Jul;35(4):397-401. [Medline: 9327619]
19. Tappan J, Daniels J, Slavin B, Lim J, Brant R, Ansermino J. Visual cueing with context relevant information for reducing change blindness. J Clin Monit Comput 2009 Aug;23(4):223-232. [doi: 10.1007/s10877-009-9186-8] [Medline: 19544053]
20. Michels P, Gravenstein D, Westenskow DR. An integrated graphic data display improves detection and identification of critical events during anesthesia. J Clin Monit 1997 Jul;13(4):249-259. [Medline: 9269619]
21. Effken J, Kim N, Shaw R. Making the constraints visible: testing the ecological approach to interface design. Ergonomics 1997 Jan;40(1):1-27. [doi: 10.1080/001401397188341] [Medline: 8995046]
22. Görges M, Westenskow DR, Markewitz BA. Evaluation of an integrated intensive care unit monitoring display by critical care fellow physicians. J Clin Monit Comput 2012 Dec;26(6):429-436. [doi: 10.1007/s10877-012-9370-0] [Medline: 22588528]
23. Stylianides N, Dikaiakos M, Gjermundrød H, Panayi G, Kyprianou T. Intensive care window: real-time monitoring and analysis in the intensive care environment. IEEE Trans Inf Technol Biomed 2011 Jan;15(1):26-32. [doi: 10.1109/TITB.2010.2091141] [Medline: 21062685]

XSL•FO
RenderX

24. Litt H, Loonsk J. Digital patient records and the medical desktop: an integrated physician workstation for medical informatics training. Proc Annu Symp Comput Appl Med Care 1992:555-559 [FREE Full text] [Medline: 1482935]
25. Gschwandtner T, Aigner W, Kaiser K, Miksch S, Seyfang A. CareCruiser: exploring and visualizing plans, events, and effects interactively. 2011 Presented at: IEEE Pacific Visualization Symposium (PacificVis); 2011; Hong Kong p. 43-50. [doi: 10.1109/PACIFICVIS.2011.5742371]
26. Horn W, Popow C, Unterasinger L. Support for fast comprehension of ICU data: visualization using metaphor graphics. Methods Inf Med 2001;40(5):421-424. [Medline: 11776741]
27. Dayhoff R, Kirin G, Pollock S, Miller C, Todd S. Medical data capture and display: the importance of clinicians' workstation design. Proc Annu Symp Comput Appl Med Care 1994:541-545 [FREE Full text] [Medline: 7949987]
28. Norris P, Dawant B. Closing the loop in ICU decision support: physiologic event detection, alerts, and documentation. Proc AMIA Symp 2001:498-502 [FREE Full text] [Medline: 11825238]
29. Langner P. The value of high fidelity electrocardiography using the cathode ray oscillograph and an expanded time scale. Circulation 1952 Feb;5(2):249-256. [Medline: 14896469]
30. Burykin A, Peck T, Krejci V, Vannucci A, Kangrga I, Buchman T. Toward optimal display of physiologic status in critical care: I. Recreating bedside displays from archived physiologic data. J Crit Care 2011 Feb;26(1):105.e1-105.e9. [doi: 10.1016/j.jcrc.2010.06.013] [Medline: 20813491]
31. Miller A, Scheinkestel C, Steele C. The effects of clinical information presentation on physicians' and nurses' decision-making in ICUs. Appl Ergon 2009 Jul;40(4):753-761. [doi: 10.1016/j.apergo.2008.07.004] [Medline: 18834970]
32. Kruger G, Tremper K. Advanced integrated real-time clinical displays. Anesthesiol Clin 2011 Sep;29(3):487-504. [doi: 10.1016/j.anclin.2011.05.004] [Medline: 21871406]
33. Law A, Freer Y, Hunter J, Logie R, McIntosh N, Quinn J. A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. J Clin Monit Comput 2005 Jun;19(3):183-194. [doi: 10.1007/s10877-005-0879-3] [Medline: 16244840]
34. 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]
35. Sainsbury D. An object-oriented approach to data display and storage: 3 years experience, 25,000 cases. Int J Clin Monit Comput 1993 Nov;10(4):225-233. [Medline: 8270836]
36. Zhang Y, Drews F, Westenskow D, Foresti S, Agutter J, Bermudez J, et al. Effects of integrated graphical displays on situation awareness in anaesthesiology. Cogn Technol Work 2002 Jun 1;4(2):82-90. [doi: 10.1007/s101110200007]
37. Kennedy RR, Merry AF, Warman GR, Webster CS. The influence of various graphical and numeric trend display formats on the detection of simulated changes. Anaesthesia 2009 Nov;64(11):1186-1191 [FREE Full text] [doi: 10.1111/j.1365-2044.2009.06082.x] [Medline: 19825052]
38. Lowe A, Jones R, Harrison M. The graphical presentation of decision support information in an intelligent anaesthesia monitor. Artif Intell Med 2001;22(2):91. [Medline: 11348846]
39. Charbonnier S. On line extraction of temporal episodes from ICU high-frequency data: a visual support for signal interpretation. Comput Methods Programs Biomed 2005 May;78(2):115-132. [doi: 10.1016/j.cmpb.2005.01.003] [Medline: 15848267]
40. Shabot M, Carlton P, Sadoff S, Nolan-Avila L. Graphical reports and displays for complex ICU data: a new, flexible and configurable method. Comput Methods Programs Biomed 1986 Mar;22(1):111-116. [Medline: 3634666]
41. Douglas JR, Ritter MJ. Implementation of an Anesthesia Information Management System (AIMS). Ochsner J 2011;11(2):102-114 [FREE Full text] [Medline: 21734847]
42. Koch S, Weir C, Westenskow D, Gondan M, Agutter J, Haar M, et al. Evaluation of the effect of information integration in displays for ICU nurses on situation awareness and task completion time: a prospective randomized controlled study. Int J Med Inform 2013 Aug;82(8):665-675. [doi: 10.1016/j.ijmedinf.2012.10.002] [Medline: 23357614]
43. Kamaleswaran R, Collins C, James A, McGregor C. PhysioEx: visual analysis of physiological event streams. Comput Graphics Forum 2016 Jul 04;35(3):331-340. [doi: 10.1111/cgf.12909]
44. Wachter S, Agutter J, Syroid N, Drews F, Weinger M, Westenskow D. The employment of an iterative design process to develop a pulmonary graphical display. J Am Med Inform Assoc 2003;10(4):363-372 [FREE Full text] [doi: 10.1197/jamia.M1207] [Medline: 12668693]
45. Anders S, Albert R, Miller A, Weinger MB, Doig AK, Behrens M, et al. Evaluation of an integrated graphical display to promote acute change detection in ICU patients. Int J Med Inform 2012 Dec;81(12):842-851 [FREE Full text] [doi: 10.1016/j.ijmedinf.2012.04.004] [Medline: 22534099]
46. Kilman D, Forslund D. An international collaboratory based on virtual patient records. Commun ACM 1997;40(8):110-117. [doi: 10.1145/257874.257898]
47. van Bemmel J, van Ginneken A, Stam B, van Mulligen E. Virtual electronic patient records for shared care. Stud Health Technol Inform 1998;52:37-41. [Medline: 10384551]
48. Tange H. The paper-based patient record: is it really so bad? Comput Methods Programs Biomed 1995;48(1-2):127-131. [Medline: 8846696]

XSL•FO
RenderX

49. Margulies D, McCallie J, Elkowitz A, Ribitzky R. An integrated hospital information system at children's hospital. Presented at: Proceedings of the Annual Symposium on Computer Application in Medical Care; 1990; New York p. 699.
50. Cole W, Stewart J. Metaphor graphics to support integrated decision making with respiratory data. Int J Clin Monit Comput 1993 May;10(2):91-100. [Medline: 8366316]
51. Were M, Shen C, Bwana M, Emenyonu N, Musinguzi N, Nkuyahaga F, et al. Creation and evaluation of EMR-based paper clinical summaries to support HIV-care in Uganda, Africa. Int J Med Inform 2010 Feb;79(2):90-96 [FREE Full text] [doi: 10.1016/j.ijmedinf.2009.11.006] [Medline: 20036193]
52. Miller A, Sanderson P. Evaluating an information display for clinical decision making in the intensive care unit. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2003 Oct 01;47(3):576-580. [doi: 10.1177/154193120304700368]
53. Gather U, Imhoff M, Fried R. Graphical models for multivariate time series from intensive care monitoring. Stat Med 2002 Sep 30;21(18):2685-2701. [doi: 10.1002/sim.1209] [Medline: 12228885]
54. Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. 1956. Psychol Rev 1994 Apr;101(2):343-352. [Medline: 8022966]
55. Eppler M, Mengis J. The concept of information overload: a review of literature from organization science, accounting, marketing, MIS, and related disciplines. Inf Soc 2004 Nov;20(5):325-344 [FREE Full text] [doi: 10.1080/01972240490507974]
56. de Oliveira M, Levkowitz H. From visual data exploration to visual data mining: a survey. IEEE Trans Visual Comput Graphics 2003 Jul;9(3):378-394 [FREE Full text] [doi: 10.1109/TVCG.2003.1207445]
57. Bawden D, Robinson L. The dark side of information: overload, anxiety and other paradoxes and pathologies. J Inf Sci 2008 Nov 21;35(2):180-191 [FREE Full text] [doi: 10.1177/0165551508095781]
58. Speier C, Valacich J, Vessey I. The influence of task interruption on individual decision making: an information overload perspective. Decis Sci 1999 Mar;30(2):337-360 [FREE Full text] [doi: 10.1111/j.1540-5915.1999.tb01613.x]
59. Endsley M. Toward a theory of situation awareness in dynamic systems. Hum Factors 1995 Mar 01;37(1):32-64 [FREE Full text] [doi: 10.1518/001872095779049543]
60. Bui A, Aberle D, Kangarloo H. TimeLine: visualizing integrated patient records. IEEE Trans Inform Technol Biomed 2007 Jul;11(4):462-473 [FREE Full text] [doi: 10.1109/TITB.2006.884365]
61. Duncan R, Saperia D, Dulbandzhyan R, Shabot M, Polaschek J, Jones D. Integrated Web-based viewing and secure remote access to a clinical data repository and diverse clinical systems. Proc AMIA Symp 2001:149-153 [FREE Full text] [Medline: 11825172]
62. Meyer M, Levine W, Brzezinski P, Robbins J, Lai F, Spitz G, et al. Integration of hospital information systems, operative and peri-operative information systems, and operative equipment into a single information display. AMIA Annu Symp Proc 2005:1054 [FREE Full text] [Medline: 16779341]
63. Sanderson P. The multimodal world of medical monitoring displays. Appl Ergon 2006 Jul;37(4):501-512. [doi: 10.1016/j.apergo.2006.04.022] [Medline: 16759627]
64. Georgopoulos D, Prinianakis G, Kondili E. Bedside waveforms interpretation as a tool to identify patient-ventilator asynchronies. Intensive Care Med 2006 Jan;32(1):34-47. [doi: 10.1007/s00134-005-2828-5] [Medline: 16283171]
65. Enison E, Dayhoff R, Fletcher R. Graphical electrocardiogram waveforms as part of an integrated hospital system's patient record. Proc Annu Symp Comput Appl Med Care 1993:373-375 [FREE Full text] [Medline: 8130498]
66. Burns C. Putting it all together: improving display integration in ecological displays. Hum Factors 2000 Jun 01;42(2):226-241. [doi: 10.1518/001872000779656471] [Medline: 10]
67. Zinser K, Frischenschlager F. Multimedia's push into power. IEEE Spectr 1994 Jul;31(7):44-48. [doi: 10.1109/6.294947]
68. Effken J, Loeb R, Kang Y, Lin Z. Clinical information displays to improve ICU outcomes. Int J Med Inform 2008 Nov;77(11):765-777. [doi: 10.1016/j.ijmedinf.2008.05.004] [Medline: 18639487]
69. Blike GT, Surgenor SD, Whalen K, Jensen J. Specific elements of a new hemodynamics display improves the performance of anesthesiologists. J Clin Monit Comput 2000;16(7):485-491. [Medline: 12580206]
70. Görges M, Förger K, Westenskow D. A trend based decision support system for anesthesiologists improves diagnosis speed and accuracy. Presented at: Proceedings of the Annual Mountain West Biomedical Engineering Conference; 2006; Snowbird, UT.
71. Albert R, Agutter J, Syroid N, Johnson K, Loeb R, Westenskow D. A simulation-based evaluation of a graphic cardiovascular display. Anesth Analg 2007 Nov;105(5):1303-11, table of contents. [doi: 10.1213/01.ane.0000282823.76059.ca] [Medline: 17959959]
72. Havre S, Hetzler B, Nowell L. ThemeRiver: visualizing theme changes over time. IEEE Symposium on Information Visualization. InfoVis 2000:115 [FREE Full text] [doi: 10.1109/INFVIS.2000.885098]
73. Law A, Freer Y, Hunter J, Logie R, McIntosh N, Quinn J. A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. J Clin Monitoring Computing 2005;19(3):94. [doi: 10.1007/s10877-005-0879-3]
74. Drews F, Westenskow D. The right picture is worth a thousand numbers: data displays in anesthesia. Hum Factors 2006;48(1):59-71. [Medline: 16696257]

XSL•FO
RenderX

Abbreviations

0: No Change
App: Application
C: Curves
Des: Design
Eval: Evaluation
Exp: Experiment
G: Glyph
MT: Metaphoric display
Neg: Negative
NI: Not included
O: Object
OB: Object-based display
Pos: Positive
Sim: Simulated
T: Text
TB: Tabular display
WF: waveform display

Edited by CL Parra-Calderón; submitted 30.09.15; peer-reviewed by S Van de Velde, N Farre; comments to author 29.11.15; revised version received 23.02.16; accepted 06.09.16; published 21.11.16.

Please cite as:
Kamaleswaran R, McGregor C
A Review of Visual Representations of Physiologic Data
JMIR Med Inform 2016;4(4):e31
URL: http://medinform.jmir.org/2016/4/e31/
doi:10.2196/medinform.5186
PMID:27872033

©Rishikesan Kamaleswaran, Carolyn McGregor. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 21.11.2016. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the Economic Impact of Smart Care Platforms: Qualitative and Quantitative Results of a Case Study

Frederic Vannieuwenborg[1], MSc; Thomas Van der Auwermeulen[2], MSc; Jan Van Ooteghem[1], MSc; An Jacobs[2], PhD; Sofie Verbrugge[1], PhD; Didier Colle[1], PhD

[1]Internet Based Communication Networks and Services research group (IBCN), Department of Information Technology (INTEC), Ghent University - iMinds, Ghent, Belgium

[2]Studies on Media, Information & Telecommunication (SMIT), Vrije Universiteit Brussel - iMinds, Brussel, Belgium

Corresponding Author:
Frederic Vannieuwenborg, MSc
Internet Based Communication Networks and Services research group (IBCN)
Department of Information Technology (INTEC)
Ghent University - iMinds
iGent
Technologiepark Zwijnaarde 15
Ghent, 9052
Belgium
Phone: 32 93314976
Fax: 32 93314899
Email: frederic.vannieuwenborg@intec.ugent.be

Abstract

Background: In response to the increasing pressure of the societal challenge because of a graying society, a gulf of new Information and Communication Technology (ICT) supported care services (eCare) can now be noticed. Their common goal is to increase the quality of care while decreasing its costs. Smart Care Platforms (SCPs), installed in the homes of care-dependent people, foster the interoperability of these services and offer a set of eCare services that are complementary on one platform. These eCare services could not only result in more quality care for care receivers, but they also offer opportunities to care providers to optimize their processes.

Objective: The objective of the study was to identify and describe the expected added values and impacts of integrating SCPs in current home care delivery processes for all actors. In addition, the potential economic impact of SCP deployment is quantified from the perspective of home care organizations.

Methods: Semistructured and informal interviews and focus groups and cocreation workshops with service providers, managers of home care organizations, and formal and informal care providers led to the identification of added values of SCP integration. In a second step, process breakdown analyses of home care provisioning allowed defining the operational impact for home care organization. Impacts on 2 different process steps of providing home care were quantified. After modeling the investment, an economic evaluation compared the business as usual (BAU) scenario versus the integrated SCP scenario.

Results: The added value of SCP integration for all actors involved in home care was identified. Most impacts were qualitative such as increase in peace of mind, better quality of care, strengthened involvement in care provisioning, and more transparent care communication. For home care organizations, integrating SCPs could lead to a decrease of 38% of the current annual expenses for two administrative process steps namely, care rescheduling and the billing for care provisioning.

Conclusions: Although integrating SCP in home care processes could affect both the quality of life of the care receiver and informal care giver, only scarce and weak evidence was found that supports this assumption. In contrast, there exists evidence that indicates the lack of the impact on quality of life of the care receiver while it increases the cost of care provisioning. However, our cost-benefit quantification model shows that integrating SCPs in home care provisioning could lead to a considerable decrease of costs for care administrative tasks. Because of this cost decreasing impact, we believe that the integration of SCPs will be driven by home care organizations instead of the care receivers themselves.

(JMIR Med Inform 2016;4(4):e33) doi:10.2196/medinform.5012

XSL•FO
RenderX

KEYWORDS

medical informatics applications; home care services; ambulatory care information systems; evaluation studies as topic; cost-benefit analysis

Introduction

A Societal Challenge

Many parts of the world face the same social evolution: an aging society. It's a challenge because with an aging society the demand for care increases while resource availability (both human and monetary) is under pressure.

Information and Communication Technology (ICT)-enabled care services have the potential to reduce costs while maintaining or increasing the quality of care. Many examples in the primary (the general practitioners and health centers) and secondary care sectors (hospitals and specialist) already exist. Electronic health records and electronic drug prescriptions are only a couple of many examples. All these types of ICT-enabled services foster better care communication, organization, less medication or diagnostic errors, and more transparent data sharing [1].

eCare Services

In recent years, focus intensified on aging in place and how ICT-enabled services could support this. The number of ICT-supported care applications (eCare) such as remote fall detection [2], social contact enhancing applications, and telecare services to diagnose patients remotely [3] grew exponentially. This has resulted in a fragmented and scattered landscape of eCare applications. Most of the services have individual and standalone characteristics but interoperability often lacks [4].

Smart Care Platforms

In response to this barrier of noninteroperability and nonintegrated eCare services, the introduction of smart care platforms (SCPs) can be witnessed [5-9]. These SCPs allow integration, monitoring, and data exchange between a set of home care applications and services that run on a central cloud-like platform. Smart care platforms foster better care communication and information sharing among the professional, informal care providers, and care receivers [10]; therefore, SCPs are not the same as telecare services though they can support them. Furthermore, many SCPs allow the integration with various monitoring sensors that provide specific context information (eg, room temperature, movement of the person, bed detection, sound level) [11]. Longitudinal analyses of these data give meaningful insights in evolution of the condition of the care receivers and their daily life patterns. In general, the functionalities of SCPs in terms of providing services can be categorized and summarized as follows [9]:

Support Care and Care Processes

Examples of these services are: online meal delivery services, alerting specific care actors in case of certain events, care journals, and care agendas.

Sharing Care Information and Care Communication

According to the role-based rights of the involved actors (eg, GP vs informal caregiver vs care receivers), one can add, change, erase, or annotate particular information of the care receiver.

Support Social Life and Activities

Making video calls with friends or relatives or being able to share some memories with family are just some of these services that support the social life of the home care receiver.

Monitoring Services

Integration of various sensors into the homes of the care receivers allows monitoring of context data such as movement, pressure sensors to detect bed or couch presence, accelerometers to detect falls, light, noise, temperature, humidity, smoke detectors, weighing scales, and so on. Through these sensors all kinds of biometric or context information can be captured. Analysis of sensor data allows evaluations of lifestyle trends.

Most SCPs exist with one or more of the above described functionalities. In other cases, SCPs provide the basic set of functionalities, which can easily be extended by adding modular services [12]. O'CareCloudS (OCCS) [12], the SCP developed in the identically named research project is a complete cloud-based platform. The basic service set of OCCS does provide several services to foster better care information sharing and social connectivity. The complete service set covers: (1) consulting and annotating the shared care record, (2) time and task registration of the care givers, (3) care agenda and a smart task list, (4) social calendar, (5) smart messaging service, and (6) a service catalogue for additional OCCS services. In addition, modular lifestyle monitoring services can be added by installing the necessary sensors. Although SCPs can support care provisioning for all types of home care receivers, in this work focus is on elderly, as it can be expected that more elderly will stay longer at home.

Evaluating Smart Care Platforms

SCPs are believed to have a positive impact on the quality and the cost efficiency of care. But at the same time the main characteristic of SCPs, the ability to connect multiple actors, poses challenges for its adoption. Multi-actor or multi-stakeholders systems require at least a neutral and preferably a positive perceived impact for every actor involved before a successful adoption is possible [13]. Also it's not clear which actor will initiate the adoption of SCPs.

Therefore, this paper focuses on determining and quantifying the impact of integrating SCPs into present home care processes for the elderly; in other words, evaluating the potential effect or added value of SCPs.

In the literature, previous work on several aspects of the evaluation of SCPs can be found. We distinguish (1) research on the evaluation methodology and (2) results of evaluation processes of eCare services.

XSL•FO
RenderX

The nature of SCPs and eCare services in general requires a multi-aspect evaluation method. Evaluating these services solely based on their medical impact would be insufficient; also focusing on economic impact would be too narrow. Evaluating eCare services in their totality requires looking at them from several angels such as the economic perspective, the medical impact, the social aspect, the impact on the involved actors, legal issues, and technical barriers [13,14].

In this knowledge, different frameworks are developed especially for evaluating eCare services [13-16]. All of them present a model or framework that takes into account several perspectives of the impact of integrating eCare or SCP services. Salvi et al [17] presents an overall evaluation framework based on quality of eCare services in the context of ambient assisted living. This incorporates many quality characteristics such as functionality, reliability, efficiency, and usability. However, the framework does not take the economic perspective into account.

In addition to the literature on the methodologies used for evaluating eCare services or SCPs, previous work on the impact of the integration and adoption of SCPs is also available.

Bossen et al [9] conclude that integrating SCPs in the home environment of care receivers can facilitate and augment the current home care processes and enhance the cooperation between the several involved actors even more. Although larger pilot tests are needed to further evaluate the CareCoor system, initial tests revealed promising results and positive impacts for the care network.

In contrast with the results of Bossen et al, findings from the Whole Systems Demonstrator cluster randomized trials indicate that the effects of “second-generation” telecare are very limited and even without significant impact [18-20]. Except for a small benefit on psychological outcomes, the gain in quality of life is very small [18]. This also results in a very high cost-effectiveness ratio, meaning that the costs needed to obtain that small increase in quality of life are very high and far above the willingness to pay for it. According to Steventon et al [19] the telecare services as implemented in the Whole System Demonstrator do not lead to significant cost reductions in the use of care services.

Contradiction of the results of these researches indicates that more research is needed to clarify the impact of ICT-supported care service. This lack of evidence is seen as one of the barriers for adoption of eCare services [21]. The absence of the proof of positive effects also impacts the formulation of policies or new reimbursement systems [22]. This can affect the complete business model of the eCare service provider [10].

This paper identifies the expected impact of SCP integration for all actors involved. Via economic analyses, from the perspective of home care organizations, potential benefits and costs are compared with costs of current processes. Doing so, this research provides more clarity on viable economic business cases for SCPs.

Methods

Overview

The methodology consists of 2 phases (see Figure 1). In the first phase, all various forms of potential impact and benefits are identified. During a second phase a 4-step economic cost-benefit analysis was modeled from the perspective of home care organizations.

Phase 1: Impact Identification

First, expected fields of impact should be identified for each actor within the context of home care provisioning. The methodology known as the innovation Binder Approach [23] resulted after multiple iterations in input data from various perspectives such as technology, user or social, and business.

Additional input for this identification process resulted from workshops, focus groups, and semistructured and informal interviews with field experts such as managers and administrative staff members of home care organizations, home care providers, and technology providers. Both qualitative (eg, less anxiety, increased peace of mind, decreased burden of care) and quantitative impacts (eg, process excellence such as less administration or faster billing procedures) can be expected for the actors.

Phase 2: Cost-Benefit Analysis From the Perspective of Home Care Organizations

Step 1: Identifying the Affected Home Care Processes via Process Breakdown Analyses

Adopting SCPs will affect several processes needed to provide home care such as administration tasks, communication, and sharing information. Via a high-level home care process breakdown, home care organizations were able to locate the most resource-intensive processes that could be affected after integrating SCPs. After this step, the identified process steps, care scheduling and billing processes in this case, were further decomposed.

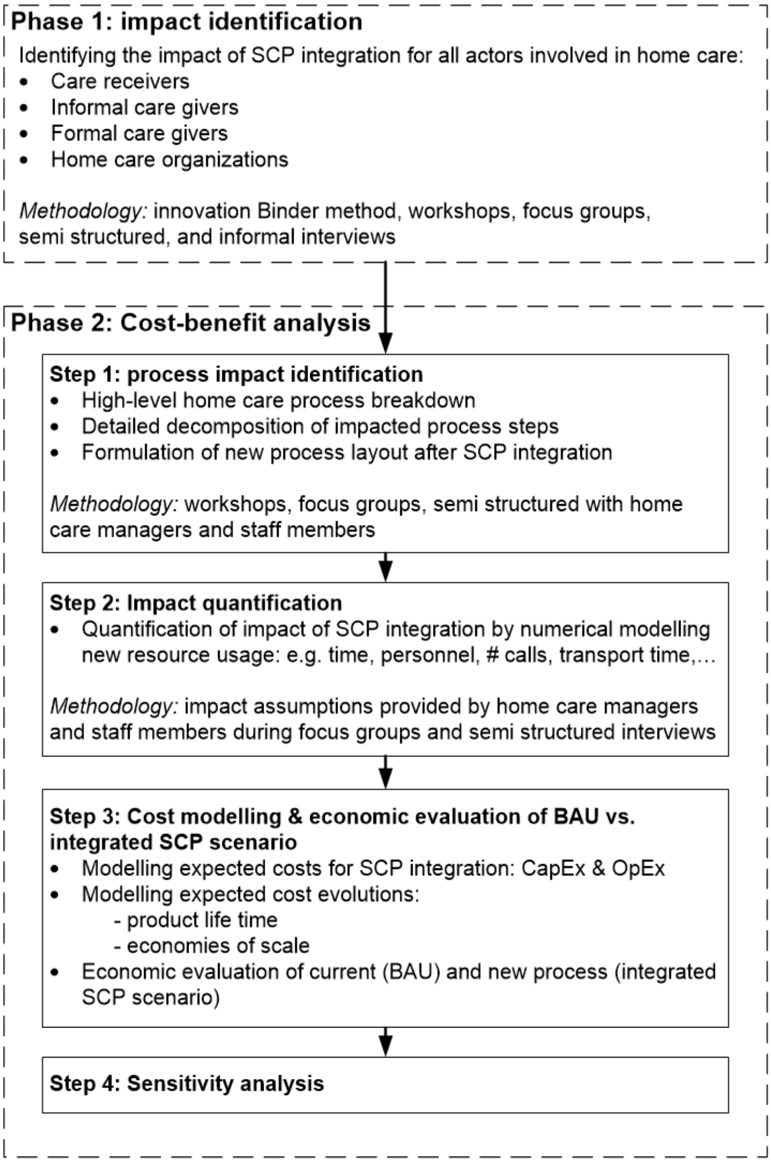
Step 2: Quantification of Costs of the Current Business as Usual and Integrated Smart Care Platform Scenario

Effects such as better scheduling and task coordination have direct quantitative impacts in terms of monetary or time savings. In this project, no qualitative or quantitative research has been carried out on the impact on health utility for care receivers such as surveying the quality of life. Therefore, this work focuses on the changes in the care scheduling and billing processes of a care organization (direct quantitative benefits).

To do so, first the annual expense of the BAU was quantified. After SCP integration, the BAU could be affected, resulting in new costs. This assumed impact, provided by home care managers and staff members during focus groups and semistructured interviews, is modeled in as well. The difference or delta between the 2 scenarios is defined as a direct benefit if the costs of the integrated SCP scenario are lower than the costs of the BAU scenario.

XSL•FO
RenderX

Figure 1. Schematic overview of the two-phase research method.


Phase 1: impact identification
Identifying the impact of SCP integration for all actors involved in home care:
• Care receivers
• Informal care givers
• Formal care givers
• Home care organizations
Methodology: innovation Binder method, workshops, focus groups, semi structured, and informal interviews
Phase 2: Cost-benefit analysis
Step 1: process impact identification
• High-level home care process breakdown
• Detailed decomposition of impacted process steps
• Formulation of new process layout after SCP integration
Methodology: workshops, focus groups, semi structured with home care managers and staff members
Step 2: Impact quantification
• Quantification of impact of SCP integration by numerical modelling new resource usage: e.g. time, personnel, # calls, transport time,…
Methodology: impact assumptions provided by home care managers and staff members during focus groups and semi structured interviews
Step 3: Cost modelling & economic evaluation of BAU vs. integrated SCP scenario
• Modelling expected costs for SCP integration: CapEx & OpEx
• Modelling expected cost evolutions:
- product life time
- economies of scale
• Economic evaluation of current (BAU) and new process (integrated SCP scenario)
Step 4: Sensitivity analysis

Step 3: Economic Evaluation: Comparing the Integrated Smart Care Platform Scenario With the Business as Usual Scenario

The goal is to research whereas the resulting benefits justify all the operational costs and investments that are needed for adopting SCPs. Thus, after quantifying the expected effects, the BAU is compared with the new “Integrated SCP scenario.” Therefore, the costs of SCP integration are also modeled.

Step 4: Dealing With Uncertainty via Sensitivity Analyses

Although this cost-benefit model is developed with realistic data provided by service providers and experts from home care organizations, it is likely that variations of the values will occur. Therefore, we need to check whether the model still behaves normally with varying input values. Sensitivity analysis also provides us with a confidence interval for the result based on the input parameters modeled with known variations.

Results

Phase 1: Overview of Potential Impact Per Actor

In the first step of this research, the potential impacts or added values, resulting from the adoption of SCPs, are identified per actor involved. Methodologies used to identify the impacts are: the “Innovation Binder Approach,” as described in [23], informal and semistructured interviews with managers of care organizations, informal and professional care providers and care receivers.

Table 1 presents the various expected added values identified per actor along with the nature of impact (qualitative or quantitative). Within the context of home care, the following actors are included: (1) care recipient or patient, (2) informal care giver, (3) formal or professional care giver and home care organization, (4) care insurers or payers and society, (5) primary care, and (6) secondary and tertiary care.

Table 1. Identification of added values that can be expected per actor.

Actor	Added value description	Impact type: qualitative or quantitative
Care receiver	Control of the organization of care	Qualitative
	Strengthened involvement and empowerment	Qualitative
	Higher quality of care	Qualitative
	Higher state of peace of mind	Qualitative
	Higher state of self-management, less care dependent	Qualitative
	Lowered barriers for social contact and decrease of social isolation	Qualitative
	Better informed of existing and practical care support services	Qualitative
Informal care giver	Better care task coordination	Qualitative
	Improved quality of care or work atmosphere	Qualitative
	Less stress, less unexpected tasks, increased state of peace of mind	Qualitative
	Being better (and real time) informed	Qualitative
Formal care giver and care organization	Better care task coordination	Qualitative
	Improved quality of care or work atmosphere	Qualitative
	Less stress, less unexpected tasks, increased state of peace of mind	Qualitative
	Significant decrease in administration time (scheduling, adapting schedules, billing, etc)	Quantitative
	Reassuring care receivers when delay during care visits	Qualitative
Primary care (GPs)	Access to more complete care and context data	Qualitative
	Improved quality of care, faster and more complete diagnoses	Qualitative
	Being better (and real time) informed	Qualitative
Secondary and tertiary care	Access to more complete care and context data	Qualitative
	Being better informed	Qualitative
	Improved quality of care, faster and more complete diagnose	Qualitative
Care insurer or payer and society	More opportunities for prevention	Qualitative
	Savings because of delayed transition to care home	Quantitative
	Increase in cost-efficiency	Quantitative
	Overall higher quality of care	Qualitative
	Transition from curative to preventive care	Qualitative

Although the potential impact for every care giver is considerable, in what follows only the impact for the care organization is quantified. This actor is considered as an SCP initiator for 2 reasons:

- Several home care organizations already provide monitoring services such as personal alarm system and work with call centers. Offering SCPs toward their clients would extend the current service offers.
- SCPs have the potential to simplify and decrease the costs for organizing home care. Therefore, home care organizations have a potential incentive to adopt SCPs.

Phase 2: Cost-Benefit Analysis From the Perspective of Home Care Organizations

The home care organizations themselves are convinced that a lot of improvement is possible in the process of home care provisioning. In order to detect which process steps would be affected by SCPs, semistructured interviews and focus groups with care organizations were carried out to collect data to be able to quantify the current costs for billing and rescheduling processes.

These data served as input for a numerical model to calculate the potential benefits and costs. In what follows, all results of the 4-step model (Figure 1) are discussed.

Step 1: Process Impact Identification

In the first step, the complete process of home care provision is broken down into several main and sub process blocks. This allowed the managers and staff members of the home care organization to locate process steps that potentially would be affected when integrating SCPs.

Figure 2 presents the high - level process spider chart for home care provisioning. The main process blocks for home care provisioning are patient intake phase, preparation of the care delivery, actual care delivery, and care delivery administration.

XSL•FO RenderX

Two-process steps were identified by the expert team as potentially impacted by adopting SCP. First, the current process for billing for home care was identified and second, the process that takes place when something has to change to the actual care schedule. For instance, when a caregiver gets sick, all planned appointments need to be reallocated to other care givers. A second example provided by the expert group is: when a client visits the hospital, all planned care visits should be replaced with others, otherwise the care givers would have no work. A more detailed decomposition of both processes is shown in Figure 3.

Step 2: Quantification of Costs of the Current Business as Usual and Integrated Smart Care Platform Scenario

The Process Break Down and Resource Usage of the Current Business as Usual Scenario

In the next step each process block of the current billing process is quantified in terms of cost per year. The same is done for the rescheduling process. Relevant data in order to calculate the cost of the current processes or business as usual are presented in Table 2.

Figure 2. High-level process breakdown of home care delivery.

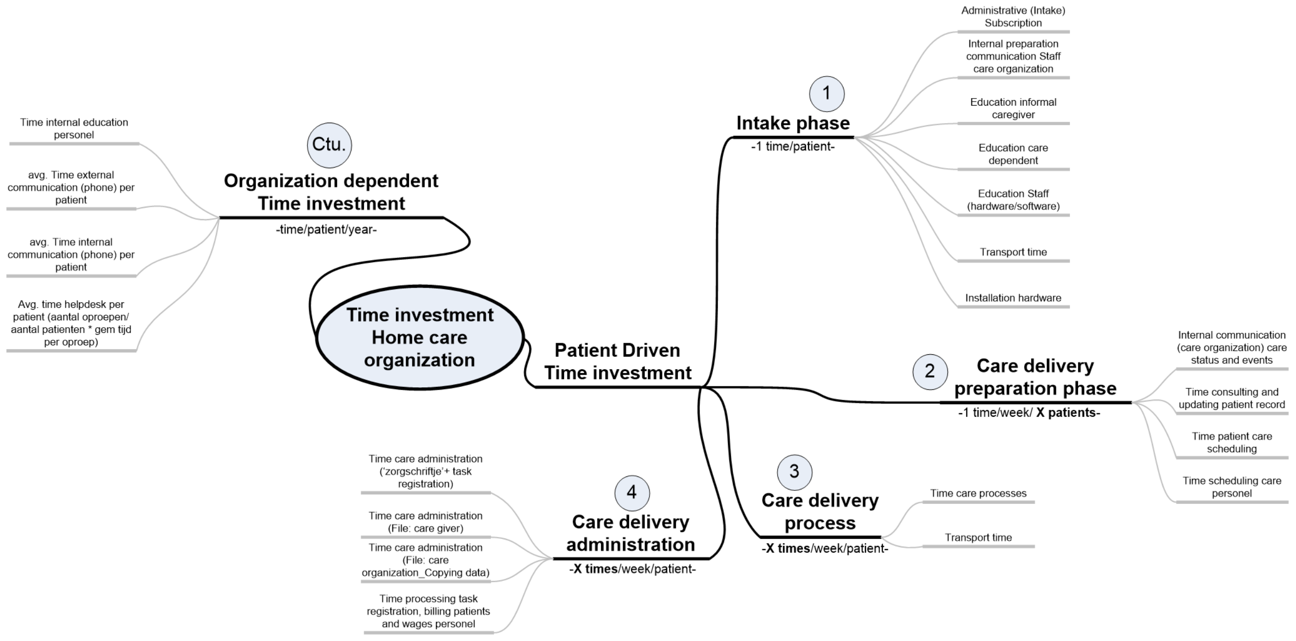

Time investment Home care organization
Ctu.
Organization dependent Time investment
-time/patient/year-
Time internal education personel
avg. Time external communication (phone) per patient
avg. Time internal communication (phone) per patient
Avg. time helpdesk per patient (aantal oproepen/ aantal patienten * gem tijd per oproep)
Patient Driven Time investment
1
Intake phase
-1 time/patient-
Administrative (Intake) Subscription
Internal preparation communication Staff care organization
Education informal caregiver
Education care dependent
Education Staff (hardware/software)
Transport time
Installation hardware
2
Care delivery preparation phase
-1 time/week/ X patients-
Internal communication (care organization) care status and events
Time consulting and updating patient record
Time patient care scheduling
Time scheduling care personel
3
Care delivery process
-X times/week/patient-
Time care processes
Transport time
4
Care delivery administration
-X times/week/patient-
Time care administration ('zorgschriftje'+ task registration)
Time care administration (File: care giver)
Time care administration (File: care organization_Copying data)
Time processing task registration, billing patients and wages personel

Chart presents the care organization blocks potentially impacted by the integration of an OCCS system.
Every block is expressed as avg. time investment per patient

XSL•FO
RenderX

Figure 3. Process decomposition of current billing and care rescheduling processes—business as usual (BAU) scenario.

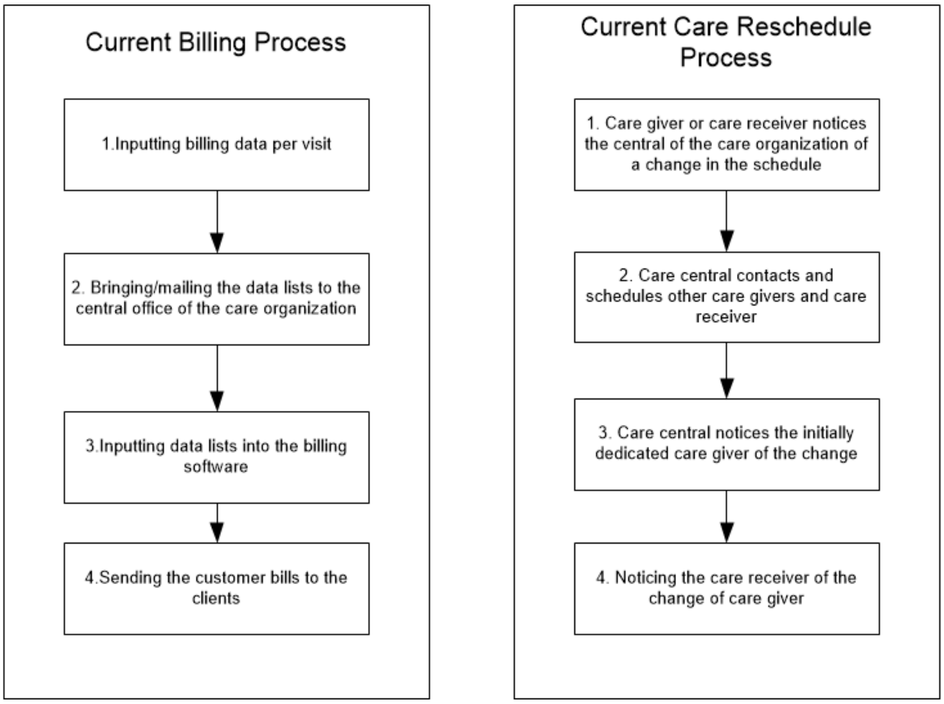
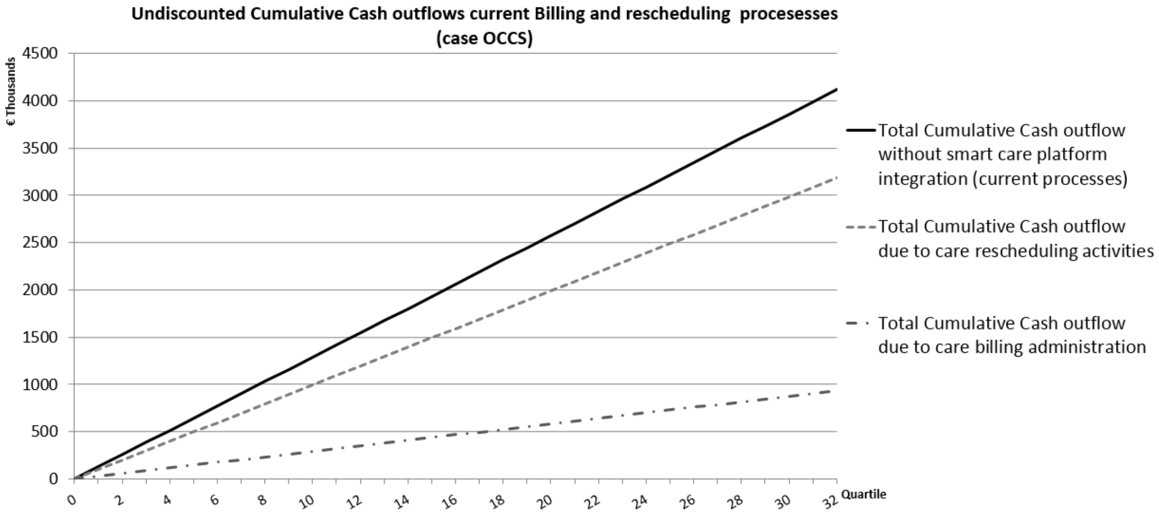

Current Billing Process
1.Inputting billing data per visit
2. Bringing/mailing the data lists to the central office of the care organization
3.Inputting data lists into the billing software
4.Sending the customer bills to the clients
Current Care Reschedule Process
1. Care giver or care receiver notices the central of the care organization of a change in the schedule
2. Care central contacts and schedules other care givers and care receiver
3. Care central notices the initially dedicated care giver of the change
4. Noticing the care receiver of the change of care giver

Table 2. Cost parameters and drivers used to calculate the cost of the business as usual (BAU) process.

Numerical parameters for the current billing process	Numerical parameters for the current rescheduling process
Number of care visits per month	Frequency of care rescheduling in terms of percentage of the total amount of planned care visits
Total amount of care givers	Telecommunication costs for calling the central administration office
Full-time equivalents (FTEs) of care providers	Average time needed to make the rescheduling exercise (not every care provider can be reallocated to a changed care visit due to professional or personal reasons (eg, care provider must speak Dutch, cannot be pregnant because of potential diseases of the cat of the care receiver)
Time needed to input the data into the back-end system	Time needed to inform the original dedicated care giver
Cost for mailing the monthly visit records of the care giver to the care organization	Scheduled visits per month
Time needed for inputting the data after each visit	Number of rescheduled visits per month
Average wages of the administration staff and the care providers	
Transport time	
Transport frequency	
Time needed for rework due to errors	

XSL•FO RenderX

Figure 4. The costs for the current rescheduling activities are more than 3 times higher than the current costs for billing administration. This is mainly caused by the wages of central office staff members who do the actual rescheduling (see Multimedia Appendix 1).



The model was initially designed for an East Flemish Care organization involved in the OCCS project, but is not limited to this organization. This region counts about 881 full-time equivalent home care givers who are members of the care organization. All data and results are valid within the scope of the OCCS project [12]. According to the managers of the care organizations, the input provided and process issues described are similar for all Flemish and even Belgian care organizations. For detailed data of current billing and rescheduling processes see Multimedia Appendix 1.

Figure 4 presents the current cumulative cash outflows per quartile for both the billing and rescheduling processes. In total, these 2 processes cost about €510,000 per year for this provincial division of the Flemish care organization involved in the OCCS project.

The Process Break Down and Resource Usage for the Integrated Smart Care Platform Scenario

Together with the care organization we modeled how an SCP would affect the current billing and rescheduling processes. Some process steps would remain unchanged; others would even disappear or would be affected. Figure 5 shows what process steps would be affected and how.

For detailed data on the affected process parameters, see Multimedia Appendix 2. Figure 6 shows the expected cumulative cash outflows of the new processes.

Given the validated impact assumptions such as reduced time needed for putting in the billing information, fewer telephone calls, no correspondence needed anymore, and so on, the total annual expense of the new processes, investments in SCP excluded would decrease to €160,000 per year. This means a reduction of 69% of the total cost of the current billing and rescheduling processes can be obtained. Figure 7 presents the comparison between the cumulative expected costs of the current and future billing and rescheduling processes.

A clear difference between the costs of the current and potential new billing and care rescheduling processes can be seen. But the latter requires a significant investment in order to reach these potential savings. Furthermore, it is expected that the data inputting process could be more time-efficient for the care provider by the use of the smart care app on the mobile phone. For the provincial home care organization involved in the OCCS project, this could free up nearly 11,000 h per year ([1488 min/year − 744 min/year] × 881 FTEs); see Multimedia Appendices 1 and for data. This time could be spent with the care receiver, resulting in better quality of care (more quality time for the patient) without affecting the cost.

XSL•FO
RenderX

Figure 5. Process decomposition of new billing and care rescheduling processes (adaptations are indicated in green)—integrated smart care platform (SCP) scenario.

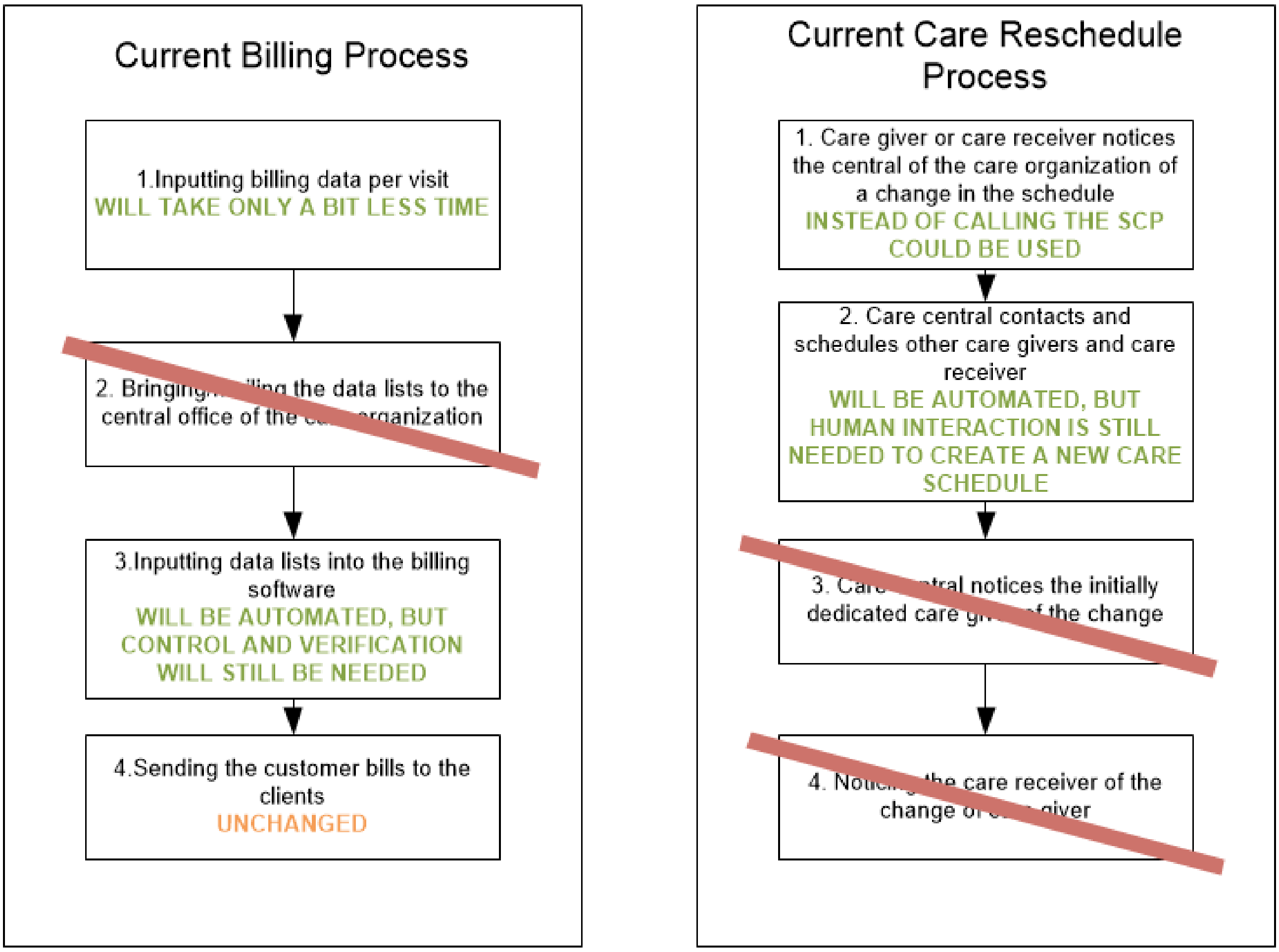
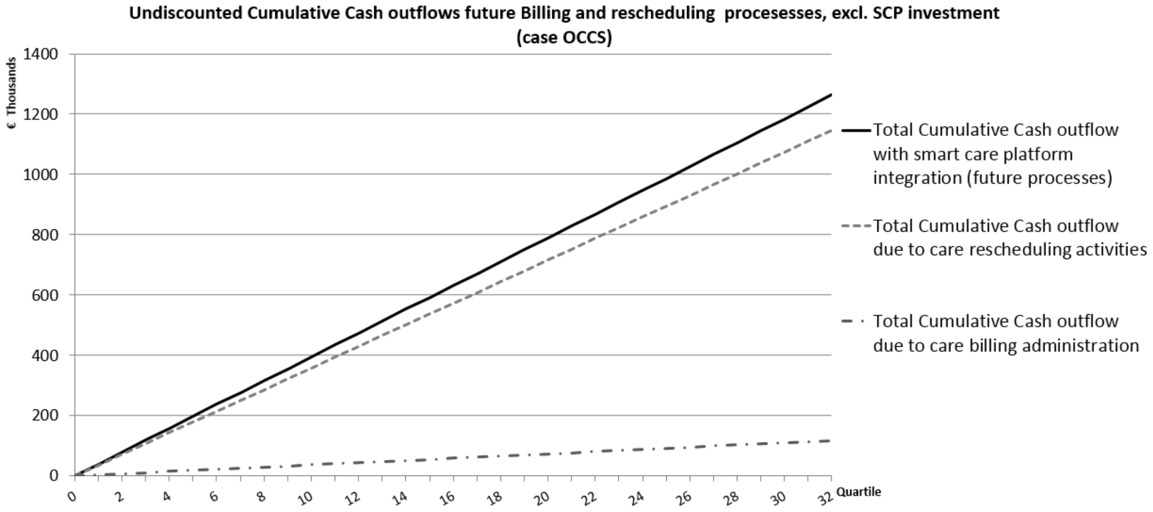

Current Billing Process
1.Inputting billing data per visit
WILL TAKE ONLY A BIT LESS TIME
~~2. Bringing ling the data lists to the central office of the rganization~~
3.Inputting data lists into the billing software
WILL BE AUTOMATED, BUT CONTROL AND VERIFICATION WILL STILL BE NEEDED
4.Sending the customer bills to the clients
UNCHANGED
Current Care Reschedule Process
1. Care giver or care receiver notices the central of the care organization of a change in the schedule
INSTEAD OF CALLING THE SCP COULD BE USED
2. Care central contacts and schedules other care givers and care receiver
WILL BE AUTOMATED, BUT HUMAN INTERACTION IS STILL NEEDED TO CREATE A NEW CARE SCHEDULE
~~3. Care tral notices the initially dedicated care g f the change~~
~~4. Noti the care receiver of the change o giver~~

Figure 6. In the new integrated smart care platform scenario, the billing process is almost completely automated. That explains the low cumulative cash outflow due to the future billing processes (see Multimedia Appendix 2).



XSL•FO
RenderX

Figure 7. Automating the billing processes and rescheduling processes would lead to a process cost reduction of 69%.

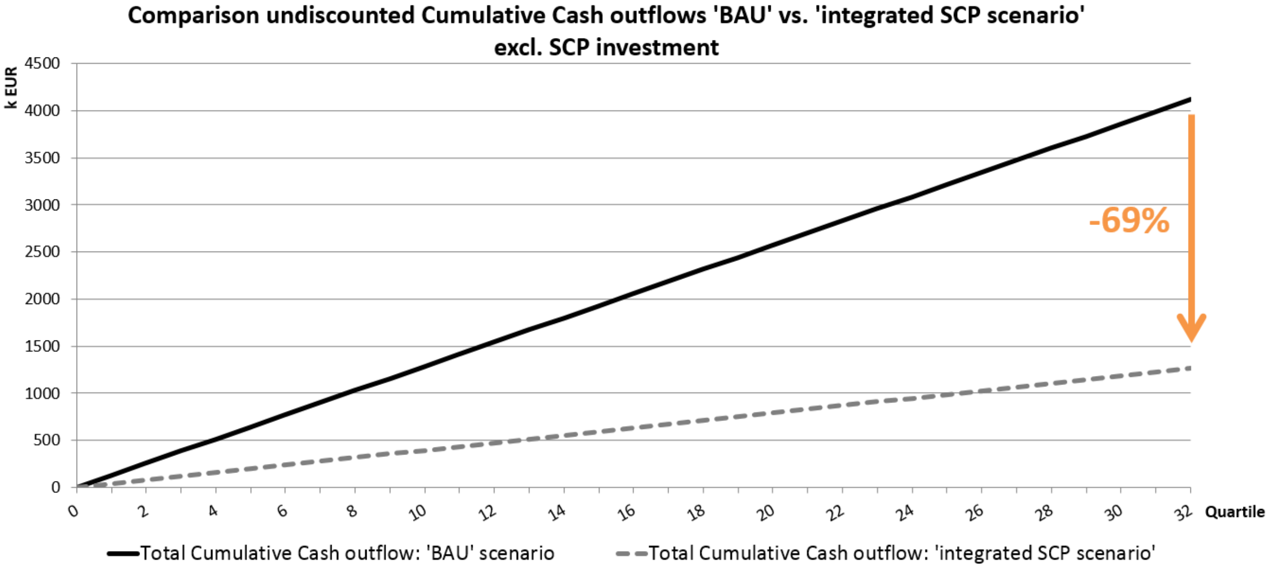


Table 3. Investments to integrate O'CareCloudS (OCCS), based on expert estimations within OCCS and sector averages.

Description of investment	Value	Unit
Every care provider needs a (basic) mobile phone, not only the people who work full-time, but also the people who work part time. (The lifetime of these devices is currently set at 3 years. Then they need to be replaced) *[CapEx]*	80	€/care provider
Every care provider needs a mobile telecommunication subscription. (There exist special group tariffs for care organization, that is why this annual expense is initially modeled rather low) *[OpEx]*	40	€/year per care provider
Each care provider needs to have access to OCCS. An annual subscription cost is modeled per care provider. *[OpEx]*	20	€/year per care provider
Each care provider needs to be educated to understand the functionalities of the SCP (2 h of education) *[CapEx]*	31	€/care provider
The SCP needs to be integrated into the back-end systems (1 FTE during 3 months) *[CapEx]*	14,700	€
An annual operational cost which is modeled as a percentage of the integration cost is needed to keep the SCP up and running *[OpEx]*	5%	

Step 3: Investment Modeling and Economic Evaluation

Smart Care Platform Investment Modeling

The expected savings can only be obtained if the home care organization invests in an SCP system like OCCS. These investments are modeled in Table 3.

Furthermore, economies of scale are modeled for the SCP subscription cost per care provider. This is modeled as a staircase function, driven by the number of care providers connected with the SCP.

The rollout of an SCP within the complete care organization is modeled as a staircase function as well. This was asked by the managers of the home care organization. Each quartile, 25% of all care givers are provided with the needed hardware and the education time. After 1 year, all care givers are connected with the SCP.

Economic Evaluation: Comparing the Integrated Smart Care Platform Scenario With the Business as Usual Scenario

Now that the impact and costs of SCP integration are known, we can investigate whether the impact is still positive after taking into account all the costs for SCP deployment.

The following graph (Figure 8) shows the expected evolution of the undiscounted cash outflow in a situation in which a smart care system would be deployed in 1 year compared with the costs of current billing and rescheduling processes.

Based on the provided data, integrating an SCP would have a payback time which is less than 1 year. Within a period of 8 years after the investment, a total cost reduction of 38% can be expected.

From Figure 9 one can see that, according to this model, the total annual expense per care provider can maximum increase to about €150 per person per year. At that level, the expected costs of the SCP integration would be the same as the current costs, everything smaller than €150 would lead to savings.

XSL•FO
RenderX

Figure 8. In the first year, the cash outflow of the integrated smart care platform (SCP) scenario is the same as for the current business as usual (BAU) scenario because of the initial investments. But after that, one can see clearly the potential savings of integrating an SCP.

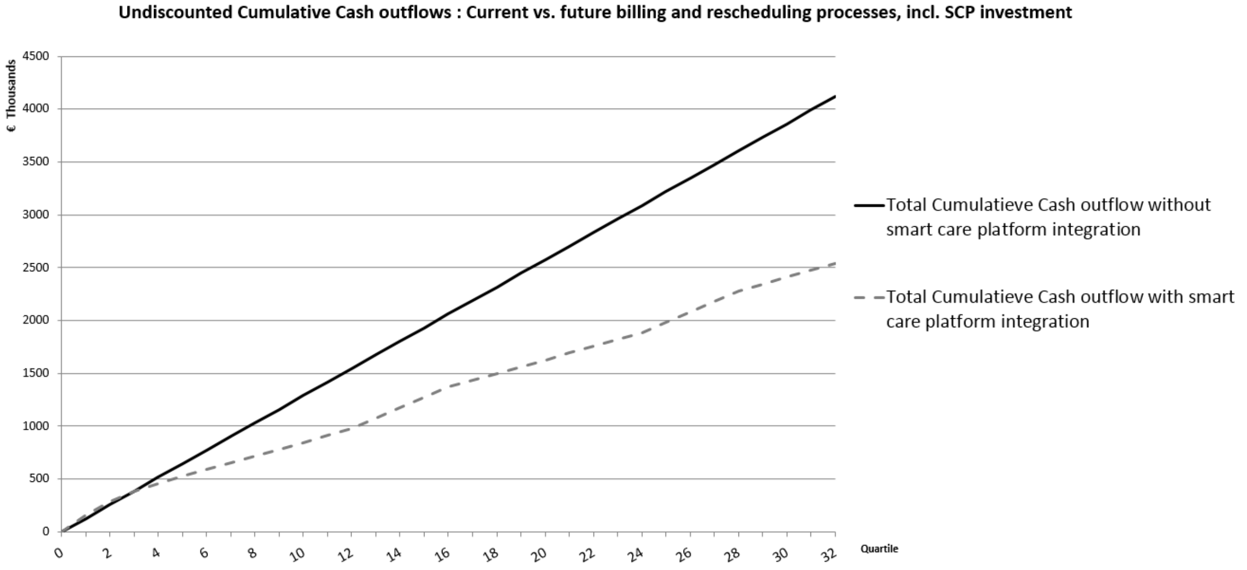


Figure 9. Expected evolution of the cumulative cash outflow in case the annual cost per care provider would be € 150. This is the upper boundary for the yearly costs per care provider.

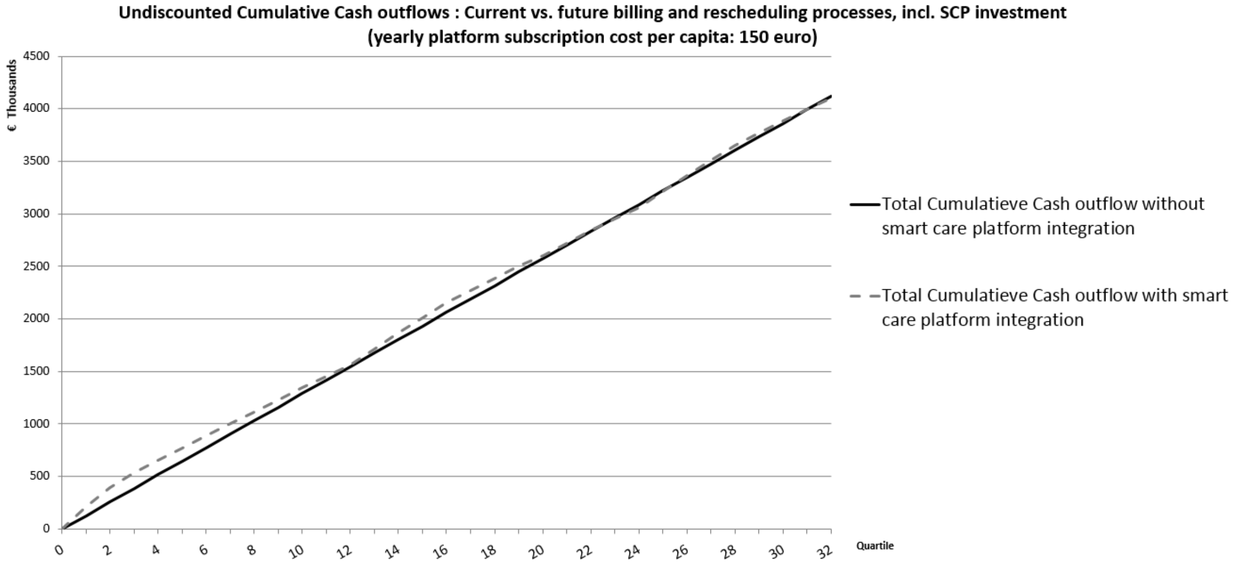


Table 4. Modeled distributions for uncertain input parameters, based on expert estimations within O'CareCloudS (OCCS).

Parameter	Modeled distribution
Number of hours needed for education (h)	Normal distribution with parameters mean=2.00, SD=0.32
Annual SCP maintenance costs (% of integration cost)	Normal distribution with parameters mean=0.05, SD=0.01
SCP back-end integration cost (€)	Lognormal distribution with parameters location=104,000, mean=14,700, SD=3498.6
Cost for mobile phone (€)	Maximum extreme distribution with parameters likeliest=80, scale=1.94
Yearly Telco subscription for the care provider (€/year)	Normal distribution with parameters mean=40, SD=11.76
Yearly smart care platform subscription cost for the care providers (€/year)	Beta distribution with parameters minimum=15, maximum=100 alpha=1.2, beta=2.6

XSL•FO
RenderX

Step 4: Dealing With Uncertainty

To take uncertainties into account, such as the assumed impacts on both affected processes, a sensitivity analysis is performed. Table 4 depicts the variations on uncertain input parameters.

The result of the sensitivity analysis indicates a 90% chance that within a period of 8 years after the investment in an SCP, the cumulative undiscounted cash outflow will lie between k€2400 and k€3400 (see Figure 10).

Testing the model robustness indicates that the SCP subscription cost is the driving parameter in this model (see Figure 11). This is acceptable as the variance on this parameter is rather high and because of the annual effect of it. The same is true for the Telco subscription cost. This means that it will be important to negotiate good subscription prices for both access to the SCP and for the telecommunication subscriptions.

Figure 10. Expected undiscounted cumulative cash outflow with CIs 90%, 50%, 25%, and 10%. In the worst-case scenario, the cost of the billing and rescheduling process will still cost 18% less than in the current situation.

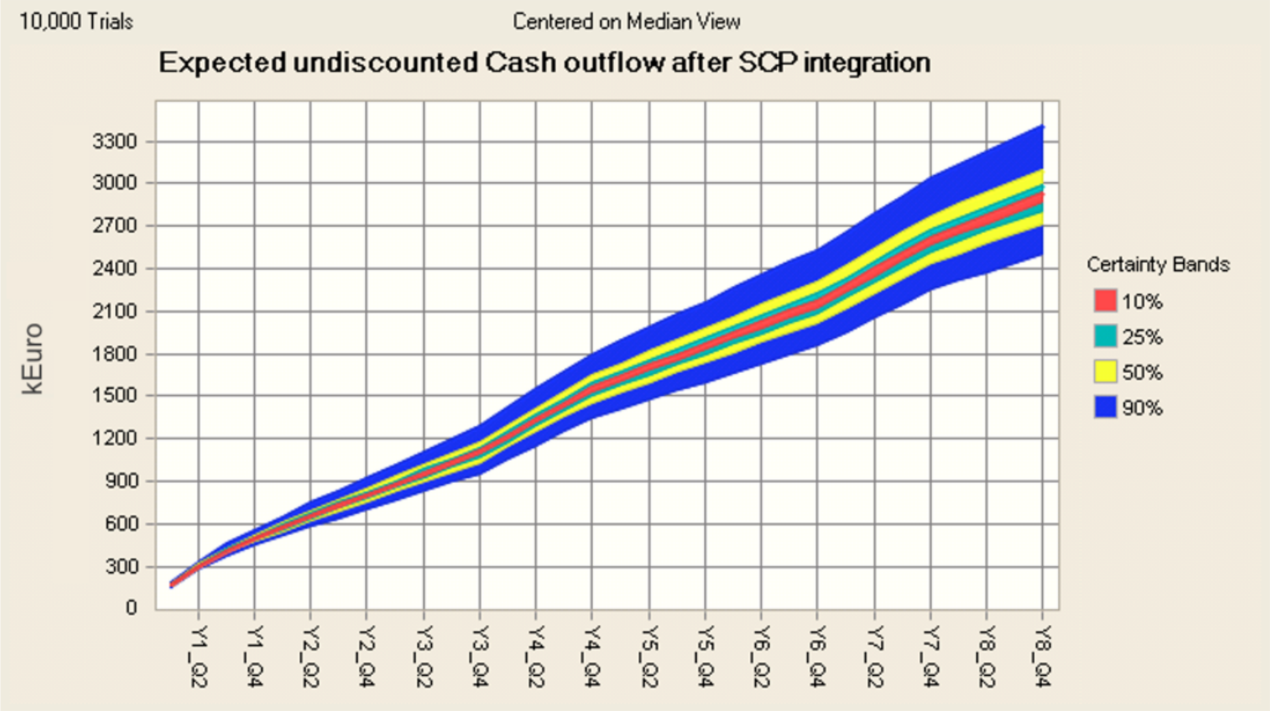
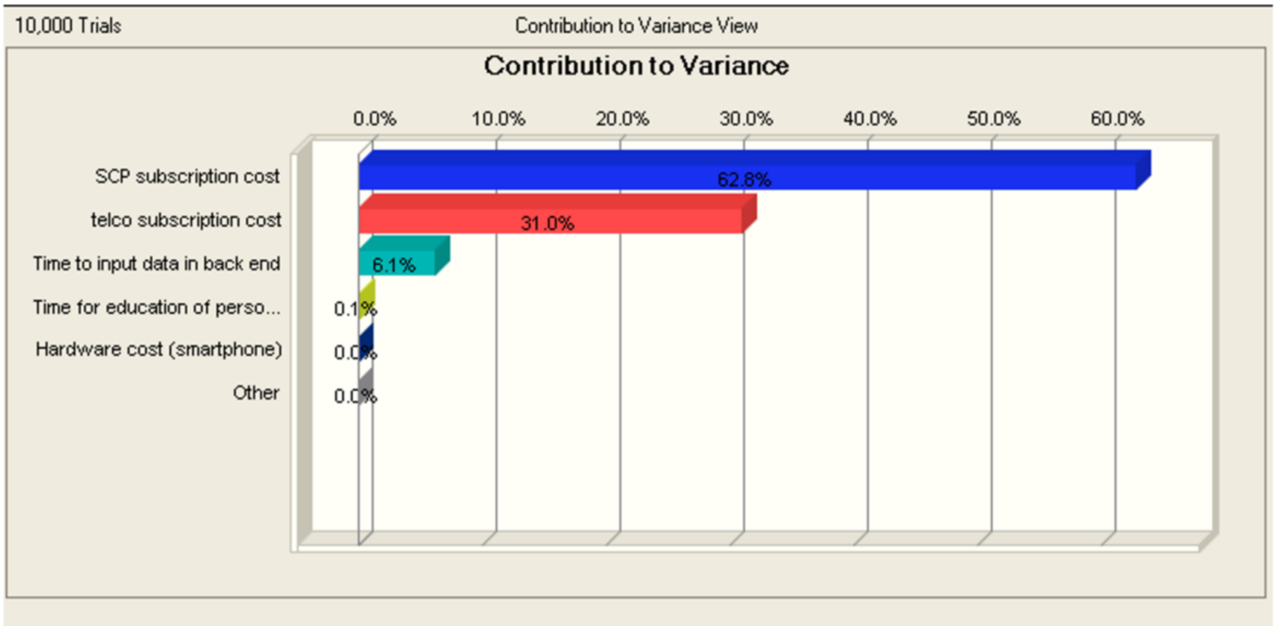


Figure 11. The annual subscription cost for smart care platforms is expected to have the biggest impact on the expected savings, followed by the annual expenses for telecommunication.



XSL•FO
RenderX

Discussion

Principal Findings

Integrating SCPs such as OCCS could affect the care administration process of care organizations. Based on the provided process data of BAU and “integrated SCP scenario,” an annual cost reduction of 37-38% could be expected. This cost reduction does not result from the SCP’s main purpose, being sharing care data or monitoring care receivers, but from the fact that digitizing one or more parts of an often time-intensive manual process can save expensive resources.

The results indicate that at least for care organizations, which are often important actors in the home care provisioning for elderly, the impact of integrating SCPs within their own scheduling and billing software is positive. This is important because literature indicates that the impact of SCPs on the quality of life of care receivers is rather limited and still not convincing enough to drive a viable adoption.

This SCP integration would not require involvement of the care receivers or their informal caregiver. Only professional care givers within the organizations could consult the care information. Initially, for the care receivers and informal care givers the added value of such a system would remain very limited.

Therefore, we believe that until the added value of SCPs for the care receiver increases to a critical level for which there exists a viable willingness to pay, the adoption of SCPs will be driven by a positive affected actor such as the care organizations. This could be a first step to digital integration and collaboration of care organizations and a first step toward a patient-centred care system.

Once all personnel of the care organization receive education and familiarize themselves with the functionalities, the home care organization can open the other functionalities of the SCPs also toward the care receivers, their informal care givers, and other care providers. In this way also other actors with a lower willingness to pay, because of the less direct quantitative benefits of SCPs, can experience the added values of SCPs.

As this research is a part of the Flemish OCCS project, the results are pertaining to the Flemish homecare organization involved in the project. As the field experts who provided data for this research stated that the situation is the same in the entire Belgium and is the same even in many Western European countries, therefore, the findings could be generalized.

Limitations

Although there are many beneficial impacts due to SCP integration, it should be noted that SCP adoption will result in some challenges and threats. Often there are concerns about privacy, data ownership, and replacing human care toward automated less personal care. It was not the focus of this research to describe all potential barriers. The results of the economic evaluation should not be affected by taking these challenges and threats into account.

Another point of remark is the single-sided perspective of the economic evaluation. Considering the case of the home care organizations alone is a well-considered choice because we strongly believe that these actors will drive adoption for SCPs. However, other actors such as society, care payer, and formal care providers could also experience economic impacts. From that point of view, the results of the analyses are probably an underestimation of the real effects. Future research on the evaluation from more perspectives complemented with a study on the impact on quality of life can bring more clarity.

Conclusions

This work envisions to identify, describe, quantify, and evaluate the impact of integrating new “Cloud-like” smart care platforms into the current home care processes. The goal of these platforms is to offer trusted information and knowledge-based services related to the care organization and delivery to the client or patient. These services aim to support and foster communication on the daily care-related needs, the social needs, and daily life assistance.

One of the goals of integrating SCPs is to foster open communication and data sharing among all the involved actors (eg, care organization, general practitioner, formal and informal care givers). Thus, in order to stimulate usage of SCPs, all actors involved should benefit from it or at least not be affected negatively.

The research indicates that all actors could benefit from the integration of SCPs. Care receivers can expect a higher quality of life, informal care givers could face a higher state of peace of mind, and formal care providers can provide the same quality of care while there could be more quality time available. Care organizations can optimize their care administration processes and push the level of digitization even further. Finally, care insurers and society in general could profit because of the possibility to provide personalized prevention and decrease or postpone the move to care homes and let the elderly stay at home instead. Although these expected effects sound acceptable, it is not clear yet whether these impacts will convince care receivers to adopt SCPs.

However, when we step away from the main goals of integrating SCPs and focus on the potential effects that result from digitizing and optimizing the current administration of home care processes (billing and care scheduling in particular), our quantification model indicates that a cost reduction for the home care organization of 37-38% could be expected and thousands of hours per year could be freed up for providing quality care by optimizing the current administrative tasks. Thus, if SCPs could be integrated within the already existing back-end systems of care organizations or vice versa, the savings potential could be a viable driver for the adoption of SCPs by home care organizations.

XSL•FO
RenderX

Acknowledgments

This research work was partially supported by the Concerted Research Actions grant for the research project: “High Potential Entrepreneurship and entrepreneurial value creation: context influences, context interactions, and context extension” (project code 01G03011) from the Ghent University and the NBB, National Bank Belgium. We acknowledge that part of this research was also supported by the iMinds Project OCCS cofunded by agentschap voor Innovatie door Wetenschap en Technologie, iMinds, and the following partners: Televic Healthcare, TPVision, Telecom IT, and Boone NV. We also thank OCMW-Kortrijk, OCMW-Ghent and Familiehulp for their cooperation in the project. The first author of this publication has also received funding from agentschap voor Innovatie door Wetenschap en Technologie.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Current process breakdown (business as usual scenario).

[PDF File (Adobe PDF File), 79KB - medinform_v4i4e33_app1.pdf]

Multimedia Appendix 2

Future process breakdown (integrated smart care platform scenario).

[PDF File (Adobe PDF File), 91KB - medinform_v4i4e33_app2.pdf]

References

1. Greenhalgh T, Stones R. Theorising big IT programmes in healthcare: strong structuration theory meets actor-network theory. Soc Sci Med 2010 May;70(9):1285-1294. [doi: 10.1016/j.socscimed.2009.12.034] [Medline: 20185218]
2. De Backere F, Ongenae F, Van den Abeele F, Nelis J, Bonte P, Clement E, et al. Social-aware and context-aware multi-sensor fall detection platform URL: http://ceur-ws.org/Vol-1114/Poster_Backere.pdf [accessed 2016-10-10] [WebCite Cache ID 6l9V1yx8z]
3. Villalba E, Casas I, Abadie F, Lluch M. Integrated personal health and care services deployment: experiences in eight European countries. Int J Med Inform 2013 Jul;82(7):626-635. [doi: 10.1016/j.ijmedinf.2013.03.002] [Medline: 23587432]
4. Czaja SJ. Long-term care services and support systems for older adults: the role of technology. Am Psychol 2016;71(4):294-301. [doi: 10.1037/a0040258] [Medline: 27159436]
5. Morris ME, Adair B, Miller K, Ozanne E, Hansen R, Pearce AJ, et al. Smart-home technologies to assist older people to live well at home. J Aging Sci 2013;1(1):1-9. [doi: 10.4172/jasc.1000101]
6. Comficare. Comficare, zelfstandig wonen in eigen hand. URL: http://www.comficare.nl/over-comficare/ [accessed 2015-05-20] [WebCite Cache ID 6YfR2wCnr]
7. Pervaya. Salveo: an intelligent telecare and home monitoring system for old persons living alone at home. URL: http://www.pervaya.com/en/index.html [accessed 2015-05-21] [WebCite Cache ID 6Yh5s7OYG]
8. Care4Balance. Care4Balance: care for balancing informal care delivery through on-demand and multi-stakeholder service design. URL: http://www.care4balance.eu/ [accessed 2015-05-21] [WebCite Cache ID 6Yh63IXIw]
9. Bossen C, Christensen LR, Grönvall E, Vestergaard LS. CareCoor: augmenting the coordination of cooperative home care work. Int J Med Inform 2013 May;82(5):e189-e199. [doi: 10.1016/j.ijmedinf.2012.10.005] [Medline: 23127539]
10. Ehrenhard M, Kijl B, Nieuwenhuis L. Market adoption barriers of multi-stakeholder technology: smart homes for the aging population.. Technological Forecasting Soc Chang 2014 Nov;89:306-315. [doi: 10.1016/j.techfore.2014.08.002]
11. Gokalp H, Clarke M. Monitoring activities of daily living of the elderly and the potential for its use in telecare and telehealth: a review. Telemed J E Health 2013 Dec;19(12):910-923. [doi: 10.1089/tmj.2013.0109] [Medline: 24102101]
12. iMinds. 2014. OCareCloudS: organizing care through trusted cloudy-like services. URL: http://www.iminds.be/en/research/overview-projects/p/detail/ocareclouds-2 [WebCite Cache ID 6Yib5Ev8U]
13. Vannieuwenborg F, van Ooteghem J, Vandenberghe M, Verbrugge S, Pickavet M, Colle D. A methodology for multi-actor evaluation of the impact of eCare services. Presented at: 2013 IEEE 15th International Conference on e-Health Networking, Applications & Services (Healthcom). 2013. p. 76-80 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6720642&isnumber=6720623 [doi: 10.1109/HealthCom.2013.6720642]
14. Kazanjian A, Green CJ. Beyond effectiveness: the evaluation of information systems using A Comprehensive Health Technology Assessment Framework. Comput Biol Med 2002 May;32(3):165-177. [Medline: 11922933]
15. van Gemert-Pijnen JE, Nijland N, van LM, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]

XSL•FO RenderX

16. Kidholm K, Ekeland AG, Jensen LK, Rasmussen J, Pedersen CD, Bowes A, et al. A model for assessment of telemedicine applications: mast. Int J Technol Assess Health Care 2012 Jan;28(1):44-51. [doi: 10.1017/S0266462311000638] [Medline: 22617736]
17. Salvi D, Montalvá Colomer JB, Arredondo MT, Prazak-Aram B, Mayer C. A framework for evaluating Ambient Assisted Living technologies and the experience of the universAAL project. AIS 2015 Jun 08;7(3):329-352. [doi: 10.3233/AIS-150317]
18. Hirani SP, Beynon M, Cartwright M, Rixon L, Doll H, Henderson C, et al. The effect of telecare on the quality of life and psychological well-being of elderly recipients of social care over a 12-month period: the Whole Systems Demonstrator cluster randomised trial. Age Ageing 2014 May;43(3):334-341 [FREE Full text] [doi: 10.1093/ageing/aft185] [Medline: 24333802]
19. Steventon A, Bardsley M, Billings J, Dixon J, Doll H, Beynon M, et al. Effect of telecare on use of health and social care services: findings from the Whole Systems Demonstrator cluster randomised trial. Age Ageing 2013 Jul;42(4):501-508 [FREE Full text] [doi: 10.1093/ageing/aft008] [Medline: 23443509]
20. Henderson C, Knapp M, Fernández J, Beecham J, Hirani SP, Beynon M, et al. Cost-effectiveness of telecare for people with social care needs: the Whole Systems Demonstrator cluster randomised trial. Age Ageing 2014 Nov;43(6):794-800 [FREE Full text] [doi: 10.1093/ageing/afu067] [Medline: 24950690]
21. Vannieuwenborg F, Kirtava Z, Lambrinos L, Van Ooteghem J, Verbrugge S. Implications of mHealth Service Deployments: A Comparison between Dissimilar European Countries. In: Stiller B, Hadjiantonis AM, editors. Telecommunication Economics. Berlin: Springer; 2012:56-66.
22. Nolte E, Emma Pitchforth E. 2014. What is the evidence on the economic impacts of integrated care? WHO Europe. URL: http://www.euro.who.int/__data/assets/pdf_file/0019/251434/What-is-the-evidence-on-the-economic-impacts-of-integrated-care.pdf [accessed 2016-10-10] [WebCite Cache ID 6l9WL9BaE]
23. Jacobs A, Duysburgh P, Bleumers L, Ongenae F, Ackaert A, Verstichel S. The innovation binder approach: A guide towards a social-technical balanced pervasive health system. In: Holzinger A, Ziefle M, Röcker C, editors. Pervasive Health: State-of-the-art and Beyond. London: Springer; 2014:69-99.

Abbreviations

BAU: business as usual
FTE: full time equivalent
OCCS: O’CareCloudS
SCP: smart care platform

Edited by G Eysenbach; submitted 07.08.15; peer-reviewed by L Simonse, F De Backere, R Drost; comments to author 17.11.15; revised version received 02.08.16; accepted 22.08.16; published 31.10.16.

Please cite as:
Vannieuwenborg F, Van der Auwermeulen T, Van Ooteghem J, Jacobs A, Verbrugge S, Colle D
Evaluating the Economic Impact of Smart Care Platforms: Qualitative and Quantitative Results of a Case Study
JMIR Med Inform 2016;4(4):e33
URL: http://medinform.jmir.org/2016/4/e33/
doi:10.2196/medinform.5012
PMID:27799137

©Frederic Vannieuwenborg, Thomas Van der Auwermeulen, Jan Van Ooteghem, An Jacobs, Sofie Verbrugge, Didier Colle. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 31.10.2016. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Email Between Patient and Provider: Assessing the Attitudes and Perspectives of 624 Primary Health Care Patients

Puneet Seth[1,2*], BSc, MD; Mohamed Ismail Abu-Abed[3*], BEng, MD; Vikram Kapoor[3*], MD; Kathryn Nicholson[4*], BHSc, MSc; Gina Agarwal[5*], MBBS, PhD

[1]Department of Family Medicine, Schulich School of Medicine & Dentistry, Western University, London, ON, Canada

[2]Division of Hospital Medicine, Woodstock General Hospital, Woodstock, ON, Canada

[3]Brampton Civic Hospital, Brampton, ON, Canada

[4]Department of Epidemiology & Biostatistics, Schulich School of Medicine & Dentistry, Western University, London, ON, Canada

[5]Family Medicine Residency Program, Department of Family Medicine, McMaster University, Hamilton, ON, Canada

[*]all authors contributed equally

Corresponding Author:
Puneet Seth, BSc, MD
Department of Family Medicine
Schulich School of Medicine & Dentistry
Western University
1151 Richmond Street
London, ON, N6A 3K7
Canada
Phone: 1 519 661 2111
Fax: 1 905 527 4440
Email: puneetsethmd@gmail.com

Abstract

Background: Email between patients and their health care providers can serve as a continuous and collaborative forum to improve access to care, enhance convenience of communication, reduce administrative costs and missed appointments, and improve satisfaction with the patient-provider relationship.

Objective: The main objective of this study was to investigate the attitudes of patients aged 16 years and older toward receiving email communication for health-related purposes from an academic inner-city family health team in Southern Ontario. In addition to exploring the proportion of patients with a functioning email address and interest in email communication with their health care provider, we also examined patient-level predictors of interest in email communication.

Methods: A cross-sectional study was conducted using a self-administered, 1-page survey of attitudes toward electronic communication for health purposes. Participants were recruited from attending patients at the McMaster Family Practice in Hamilton, Ontario, Canada. These patients were aged 16 years and older and were approached consecutively to complete the self-administered survey (N=624). Descriptive analyses were conducted using the Pearson chi-square test to examine correlations between variables. A logistic regression analysis was conducted to determine statistically significant predictors of interest in email communication (yes or no).

Results: The majority of respondents (73.2%, 457/624) reported that they would be willing to have their health care provider (from the McMaster Family Practice) contact them via email to communicate health-related information. Those respondents who checked their personal email more frequently were less likely to want to engage in this electronic communication. Among respondents who check their email less frequently (fewer than every 3 days), 46% (37/81) preferred to communicate with the McMaster Family Practice via email.

Conclusions: Online applications, including email, are emerging as a viable avenue for patient communication. With increasing utility of mobile devices in the general population, the proportion of patients interested in email communication with their health care providers may continue to increase. When following best practices and appropriate guidelines, health care providers can use this resource to enhance patient-provider communication in their clinical work, ultimately leading to improved health outcomes and satisfaction with care among their patients.

XSL•FO
RenderX

(JMIR Med Inform 2016;4(4):e42) doi:10.2196/medinform.5853

KEYWORDS

electronic mail; email; communication; primary health care; surveys; patient engagement

Introduction

The use of the Internet and electronic communication for day-to-day purposes is becoming an increasingly ubiquitous resource in many developed countries around the world [1]. The use of technology and electronics in health care delivery is also continuing to rise in prevalence [2-7]. Among other modalities [8], email between patients and their health care providers can serve as a continuous and collaborative forum to improve access to care, enhance convenience of communication outside of traditional office hours, reduce administrative costs and missed appointments, and improve satisfaction with the patient-provider relationship [2,9-14]. A systematic review conducted by Ye et al (2010) included content analyses of email messages between patients and health care providers and indicated that emails were commonly used for medical information exchange, medical condition or update, medication information, and subspecialty evaluation [12].

The benefits and risks associated with using email communication have been well-articulated in previous literature [2,6,7,9,14,15]. The potential advantages of email in delivering health care include (1) increased convenience for patients and providers (eg, time savings, avoiding need for in-person visit) [2,9-11]; (2) the continuous recording of health-related information (eg, tests results, addresses and telephone numbers of referrals, postoperative instructions) [2,10]; (3) increased opportunity for information sharing (eg, sending educational material relevant to their health) [2,10]; and (4) a user-friendly medium for patients to ask clarification questions after a face-to-face consultation [2,12]. However, there is concern from health care providers that improper use of this resource may hinder the patient-provider relationship [2,4,5], become a source of legal liability [12,15], increase the risk of diagnostic or communication errors [2,15], highlight social disparities among patients [2,14,16], and threaten patient privacy [2,4,12,15,17-19]. Providers have also been wary of adopting email as a major mode of communication with their patients, citing concerns of reimbursement, inundation with email, time demands, and the possibility of dealing with trivial issues or topics that are inappropriate to manage over email [4,17,19-21]. Despite these concerns, some studies have indicated that the email medium has promise in improving communication and access in health care. For example, patients tended to use the format appropriately by avoiding emergent issues, limiting the content to medical and administrative-oriented topics (eg, arranging appointments), and including only one request per email [9,12,22,23].

The main objective of this study, conducted as part of a Quality Assurance project at McMaster Family Practice, is to investigate the attitudes of patients aged 16 years and older toward receiving email communication for health-related purposes from an academic inner-city family health team in Southern Ontario. This was achieved through the development and distribution of a questionnaire by the study authors that identified patient concerns around email communication, their willingness to use this modality for communication from the clinic, and what specific purposes they felt would be most useful.

Methods

Setting and Study Sample

The project took place at McMaster Family Practice in Hamilton, Ontario, Canada. McMaster Family Practice is a large academic family medicine clinic situated in the downtown of an urban region that provides a full range of comprehensive primary care, with a particular focus on inner city health issues. Patients aged 16 years and older, who attended the clinic, were eligible to participate in the survey. Patient recruitment occurred during the time of checking in for a clinic visit with the medical office assistant. Patients meeting eligibility criteria (greater than 16 years of age, fluent in English, and without any diagnosis of cognitive impairment) were offered the opportunity to participate in the study. If they agreed, they were provided with a clipboard with the questionnaire and a pen—there was no digital modality offered for this questionnaire. Patients who agreed to complete the pseudonymous survey were compensated for their participation with a small treat (value less than Can $1), and completed the survey in the practice waiting room before their health care encounter. Approval for the project was granted by the Hamilton Integrated Research Ethic Board.

Study Design and Data Collection

The study was a cross-sectional, self-administered survey of patients who met the inclusion criteria at the date of data collection. The survey instrument was a 1-page, 2-sided document that was developed by the authors following a literature review and discussion (see Multimedia Appendix 1). In addition to demographic characteristics, respondents were asked about their satisfaction or dissatisfaction of the potential to use email communication with their health care provider. Responses from completed surveys were entered into an electronic database for analysis. Surveys were completed anonymously, with only their personal identifiers (the first three digits of the patient's residential postal code) and patient age at date of data collection.

Data Analysis

Descriptive analyses were conducted to examine participant characteristics, frequencies of responses, and relationships between key variables. A Pearson chi-square test was conducted to explore correlations between variables, and a stepwise logistic regression analysis was conducted to identify the independent variables that were statistically significant predictors of the dependent variable, which was patient interest in email communication (yes or no). The distribution of independent and dependent variables was explored before analysis. The significance level was set to .05, while case-wise deletion was

XSL•FO
RenderX

used for missing data. All analyses were conducted using SPSS Version 19.

Results

Participant Characteristics

A summary of all participant characteristics and demographics is presented in Table 1. Overall, 49.7% (310/624) of respondents were female and 17.6% (110/624) were between the ages of 35 and 44 years. Slightly less than half of the respondents had completed university-level education (43.1%, 269/624) and were employed at the time of the study (47.6%, 297/624). While 87.5% (546/624) of respondents stated that they had a personal email address, 73.2% (457/624) of patients stated that they would be willing to have health-related email communication with the McMaster Family Practice.

Table 1. Characteristics and demographics of study participants (N=624).

Patient Characteristics	n (%)
Sex	
Male	186 (29.8)
Female	310 (49.7)
Not specified	128 (20.5)
Age, years	
16-24	47 (7.5)
25-34	102 (16.3)
35-44	110 (17.6)
45-54	105 (16.8)
55-64	86 (13.8)
65-74	51 (8.2)
75-84	26 (4.2)
85-94	3 (0.5)
Not specified	94 (15.1)
Education level	
Less than high school	16 (2.6)
High school	75 (12.0)
College	129 (20.7)
Undergraduate	134 (21.5)
Postgraduate	135 (21.6)
Not specified	135 (21.6)
Employment status	
Employed	297 (47.6)
Retired	98 (15.7)
Unemployed	67 (10.7)
Not specified	162 (26.0)

Willingness to Use Email Communication

The correlation between how often a participant checked their email and their willingness to receive email communication was assessed and is presented in Table 2. A total of 90.6% (414/457) of respondents who checked their email frequently (at least once every 3 days) were willing to be contacted by email, as compared to 45.7% (37/81) of participants who checked their email less frequently (*P*<.001). Interestingly, the willingness to be contacted did not vary by patient age (*P*=.30) or patient sex (*P*=.95). In examining the influence of education level, 88.1% (119/135) of patients who did have a postgraduate education were open to email communication, while 77.0% (271/352) of respondents who did not have a postgraduate education were still open to email communication (*P*<.001). A total of 70.0% (437/624) of patient respondents did not have an interest in SMS text messaging (short message service, SMS) communication. This trend was evident regardless of age group. When asked about privacy concerns, 63.3% (395/624) of respondents were not concerned or only somewhat concerned. However, 24.7% (154/624) of patients stated that privacy was a serious concern and the remaining 12.0% (75/624) of respondents were unsure or undecided. Among patients that were not concerned or only somewhat concerned about privacy,

XSL•FO
RenderX

87.7% (270/308) were willing to be contacted by email, as compared to 74.2% (89/120) of respondents who were concerned or very concerned about privacy (*P*<.001).

In the logistic regression analysis, which determined predictors of respondent interest in health-related email communication, 3 independent variables were found to be statistically significant (*P*<.05). The final model is presented in Table 3. Among those patients who accepted text messages, there was a 3.7-fold increase in odds of whether these patients would also want to utilize email communication (*P*=.002), holding all other variables constant. Among patients who utilized personal email, there was an 8.3-fold increase in odds of whether the patient would also want to utilize email communication with their health care provider (*P*=.03), holding all other variables constant. Finally, patients who checked their email frequently were 58% less likely to be interested in email communication (*P*<.001), holding all other variables constant.

Table 2. Variable correlation with participant interest in health-related email communication.

Independent variable	Chi square test *P* value (degrees of freedom)
Concerned about privacy	<.001[a] (5)
Concerned about junk mail	.45 (4)
Email benefit	<.001[a] (4)
Frequently checks email	<.001[a] (4)
Forgetting appointments	.01[b] (5)
Overall satisfaction with current communication	.10 (4)

[a]Statistically significant, *P*<.01.

[b]Statistically significant, *P*<.05.

Table 3. Logistic regression analysis examining predictor variables of participant interest in health-related email communication.

Independent Variable	exp (B)[a]	*P* value[b]
Age category	1.16	.16
Sex	1.15	.67
More education	0.00	>.99
Less education	0.00	>.99
Employed	0.73	.33
Use of text messaging	3.72	<.001[c]
Frequently checks email	0.42	<.001[c]
Personal email	8.29	.03[d]
Overall satisfaction with current communication	1.00	>.99

[a]Exponentiation of the B coefficient (odds ratio).

[b]Controlling for all other independent variables in the model.

[c]Statistically significant, *P*<.01.

[d]Statistically significant, *P*<.05.

Discussion

Principal Findings

The vast majority of respondents (73.2%,457/624) reported willingness to communicate electronically with their family practice for health-related purposes, which is comparable to previous research that has found a large proportion of patients (70% to 90%) had access to email and interest in using email to communicate with their health care provider [14,16,17,20]. Increasing interest and openness to electronic communication highlights the “technological revolution” that has occurred in everyday life for patients [1]. The disinterest in text messaging and concerns regarding privacy in our survey respondents has likely lessened as a result of increased utility of mobile devices in general publication since the time of this study, as well as the improved public perception and comfort with health-related use of information technology [24]. Our study indicated that despite concern for confidentiality, 74.2% (194/334) of these patients would still allow for email communication. Those respondents who checked their personal email more frequently were also more likely to want to engage in health-related email communication. These individuals may have technology and electronic communication more fully integrated into their daily lives, such as through the use of mobile devices. However, those patients who did not have a personal email and who were not interested in engaging with their health care providers electronically represent an important demographic, who must

XSL•FO
RenderX

not be left behind in this "technological revolution." Interestingly, of the respondents who check their email less frequently than every 3 days, 45.7% (37/81) would still be interested in communicating with the McMaster Family Practice via email. This finding may indicate that patients would be interested in making use of their email for specific purposes, such as for health-related communication and decisions.

Implications for Practice, Policy, and Research

This study indicates that email communication could provide an important avenue for health-related information between interested patients and providers. When used to its highest potential, electronic communication could enhance convenience, access, information sharing, satisfaction, and quality of care. However, at its basic level, email communication can have an impact on allowing for electronic scheduling and appointment reminders, as well as the opportunity for clarification after a face-to-face encounter with a primary health care provider or a specialist. While this is an ideal outcome of this technology, it is crucial that the "technological divide" does not hinder patient experience [2,14,16]. For example, patients who do not have interest or access to regular email must be able to maintain relationships with their providers. While patients may become increasingly accepting of the use of technology in their health care encounters, regulations must be in place to ensure that confidentiality and privacy in email communication remains a priority.

Limitations

There are four key limitations to this study that have been identified. Firstly, participants in the survey were derived from a convenient sample of consecutive patients who were at least 16 years of age and who were attending the family practice on the date of data collection. While the final sample size was just over 600 patients, future studies should randomly select and survey members of the general and patient population. Secondly, those patients who did not use email at all (eg, patients who may be older, in poor health or of lower socioeconomic status) may have been less inclined to participate and complete the survey, or may not have been part of the population able to attend the clinic. As such, selection bias may have occurred in data collection and may have influenced the findings of this study. Thirdly, the patient characteristics of this family practice, which is an academic practice in an urban setting, should be considered when generalizing the results of this study. This study provides pertinent information for email communication at the McMaster Family Practice and the generalizability of these findings to other contexts or populations should be carefully assessed. Finally, opportunities for improvement of the study questionnaire itself have been identified, including the use of a scaled response grade for the questions asking about willingness to receive email and text communication from the clinic (as opposed to the dichotomous "yes" and "no" used), and asking about access to mobile devices and Internet.

Conclusions

Our survey found that the majority of participating patients have a functioning email address and are willing to use email for health-related communication with the McMaster Family Practice. The willingness to receive email communication was not significantly correlated with age, indicating that older patients were still interested in this health communication approach. Surprisingly, privacy was not a significant concern for many patients, despite privacy being a common potential issue discussed in previous literature. A wealth of research has demonstrated that effective communication between patients and providers may positively influence patient's behaviors and well-being, including satisfaction with care, medication adherence, recall and comprehension of medical information, and even functional and physiological status [25-27]. Email communication between patients and their health care providers can serve as a viable resource to enhance dialogue both inside and outside of the clinic room. While research has shown that clinicians find email communication with patients useful for administrative purposes (eg, appointment bookings, invitations, and reminders for preventive care), future research should examine whether specific approaches (eg, integration into personal health record) would make email communication more desirable for patients and providers. Future research should also assess the influence of email communication on specific aspects of the patient-provider relationship (eg, patient literacy, shared decision-making) and the best practices to maximize the effectiveness and quality of email communication between patients and their health care providers.

Acknowledgments

We thank the Department of Family Medicine at McMaster University for providing funds to purchase treats for the study participants (less than Can $200 total).

Authors' Contributions

Mohamed Ismail Abu-Abed, Vikram Kapoor, Puneet Seth, and Gina Agarwal contributed to the study concept, design, and data analysis plan. Data analysis was conducted by Gina Agarwal. Kathryn Nicholson, Puneet Seth, and Gina Agarwal drafted the manuscript. All authors contributed to the critical revision of the final manuscript and approved the final version submitted for publication.

Conflicts of Interest

Dr Puneet Seth is Chief Medical Officer at InputHealth Systems Inc, a Canadian health informatics company. The study predates his involvement in the company, and there is no financial association or link otherwise with the company. The remaining authors have no conflicts of interest to declare.

XSL•FO
RenderX

Multimedia Appendix 1

Survey instrument administered to primary health care patients.

[PDF File (Adobe PDF File), 282KB - medinform_v4i4e42_app1.pdf]

References

1. Horwitz LI, Detsky AS. Physician communication in the 21st century: to talk or to text? JAMA 2011 Mar 16;305(11):1128-1129. [doi: 10.1001/jama.2011.324] [Medline: 21406650]
2. Car J, Sheikh A. Email consultations in health care: 1--scope and effectiveness. BMJ 2004 Aug 21;329(7463):435-438 [FREE Full text] [doi: 10.1136/bmj.329.7463.435] [Medline: 15321902]
3. Liss DT, Reid RJ, Grembowski D, Rutter CM, Ross TR, Fishman PA. Changes in office visit use associated with electronic messaging and telephone encounters among patients with diabetes in the PCMH. Ann Fam Med 2014 Jul;12(4):338-343 [FREE Full text] [doi: 10.1370/afm.1642] [Medline: 25024242]
4. Patt MR, Houston TK, Jenckes MW, Sands DZ, Ford DE. Doctors who are using e-mail with their patients: a qualitative exploration. J Med Internet Res 2003;5(2):e9 [FREE Full text] [doi: 10.2196/jmir.5.2.e9] [Medline: 12857665]
5. Brooks RG, Menachemi N. Physicians' use of email with patients: factors influencing electronic communication and adherence to best practices. J Med Internet Res 2006 Mar 24;8(1):e2 [FREE Full text] [doi: 10.2196/jmir.8.1.e2] [Medline: 16585026]
6. Menachemi N, Prickett CT, Brooks RG. The use of physician-patient email: a follow-up examination of adoption and best-practice adherence 2005-2008. J Med Internet Res 2011 Feb 25;13(1):e23 [FREE Full text] [doi: 10.2196/jmir.1578] [Medline: 21447468]
7. Goodyear-Smith F, Wearn A, Everts H, Huggard P, Halliwell J. Pandora's electronic box: GPs reflect upon email communication with their patients. Inform Prim Care 2005;13(3):195-202 [FREE Full text] [Medline: 16259859]
8. de Lusignan S, Mold F, Sheikh A, Majeed A, Wyatt JC, Quinn T, et al. Patients' online access to their electronic health records and linked online services: a systematic interpretative review. BMJ Open 2014 Sep 08;4(9):e006021 [FREE Full text] [doi: 10.1136/bmjopen-2014-006021] [Medline: 25200561]
9. Leong SL, Gingrich D, Lewis PR, Mauger DT, George JH. Enhancing doctor-patient communication using email: a pilot study. J Am Board Fam Pract 2005;18(3):180-188 [FREE Full text] [Medline: 15879565]
10. Plener I, Hayward A, Saibil F. E-mail communication in the management of gastroenterology patients: a review. Can J Gastroenterol Hepatol 2014 Mar;28(3):161-165 [FREE Full text] [Medline: 24619639]
11. Wallwiener M, Wallwiener CW, Kansy JK, Seeger H, Rajab TK. Impact of electronic messaging on the patient-physician interaction. J Telemed Telecare 2009;15(5):243-250. [doi: 10.1258/jtt.2009.090111] [Medline: 19590030]
12. Ye J, Rust G, Fry-Johnson Y, Strothers H. E-mail in patient-provider communication: a systematic review. Patient Educ Couns 2010 Aug;80(2):266-273 [FREE Full text] [doi: 10.1016/j.pec.2009.09.038] [Medline: 19914022]
13. de Jong CC, Ros WJ, Schrijvers G. The effects on health behavior and health outcomes of Internet-based asynchronous communication between health providers and patients with a chronic condition: a systematic review. J Med Internet Res 2014 Jan 16;16(1):e19 [FREE Full text] [doi: 10.2196/jmir.3000] [Medline: 24434570]
14. Virji A, Yarnall KS, Krause KM, Pollak KI, Scannell MA, Gradison M, et al. Use of email in a family practice setting: opportunities and challenges in patient- and physician-initiated communication. BMC Med 2006 Aug 15;4:18 [FREE Full text] [doi: 10.1186/1741-7015-4-18] [Medline: 16911780]
15. Car J, Sheikh A. Email consultations in health care: 2--acceptability and safe application. BMJ 2004 Aug 21;329(7463):439-442 [FREE Full text] [doi: 10.1136/bmj.329.7463.439] [Medline: 15321903]
16. Moyer CA, Stern DT, Dobias KS, Cox DT, Katz SJ. Bridging the electronic divide: patient and provider perspectives on e-mail communication in primary care. Am J Manag Care 2002 May;8(5):427-433 [FREE Full text] [Medline: 12019595]
17. Kleiner KD, Akers R, Burke BL, Werner EJ. Parent and physician attitudes regarding electronic communication in pediatric practices. Pediatrics 2002 May;109(5):740-744. [Medline: 11986430]
18. Canadian Medical Protective Association. Using email communication with your patients: legal risks. 2015. URL: https://www.cmpa-acpm.ca/-/using-email-communication-with-your-patients-legal-ris-1 [accessed 2016-12-17] [WebCite Cache ID 6mpCPXI6W]
19. Hobbs J, Wald J, Jagannath YS, Kittler A, Pizziferri L, Volk LA, et al. Opportunities to enhance patient and physician e-mail contact. Int J Med Inform 2003 Apr;70(1):1-9. [Medline: 12706177]
20. Couchman GR, Forjuoh SN, Rascoe TG. E-mail communications in family practice: what do patients expect? J Fam Pract 2001 May;50(5):414-418. [Medline: 11350705]
21. Sittig DF, King S, Hazlehurst BL. A survey of patient-provider e-mail communication: what do patients think? Int J Med Inform 2001 Apr;61(1):71-80. [Medline: 11248604]
22. Anand SG, Feldman MJ, Geller DS, Bisbee A, Bauchner H. A content analysis of e-mail communication between primary care providers and parents. Pediatrics 2005 May;115(5):1283-1288. [doi: 10.1542/peds.2004-1297] [Medline: 15867036]

XSL•FO
RenderX

23. White CB, Moyer CA, Stern DT, Katz SJ. A content analysis of e-mail communication between patients and their providers: patients get the message. J Am Med Inform Assoc 2004;11(4):260-267 [FREE Full text] [doi: 10.1197/jamia.M1445] [Medline: 15064295]
24. Milward J, Day E, Wadsworth E, Strang J, Lynskey M. Mobile phone ownership, usage and readiness to use by patients in drug treatment. Drug Alcohol Depend 2015 Jan 01;146:111-115. [doi: 10.1016/j.drugalcdep.2014.11.001] [Medline: 25468818]
25. Stewart MA. Effective physician-patient communication and health outcomes: a review. CMAJ 1995 May 01;152(9):1423-1433 [FREE Full text] [Medline: 7728691]
26. Kelley JM, Kraft-Todd G, Schapira L, Kossowsky J, Riess H. The influence of the patient-clinician relationship on healthcare outcomes: a systematic review and meta-analysis of randomized controlled trials. PLoS One 2014;9(4):e94207 [FREE Full text] [doi: 10.1371/journal.pone.0094207] [Medline: 24718585]
27. Stewart M, Brown J, Weston W, McWhinney I, McWilliam C, Freeman T. Patient-Centered Medicine: Transforming the Clinical Method. Third edition. London, England: Radcliffe Publishing Ltd; 2014.

Edited by G Eysenbach; submitted 11.04.16; peer-reviewed by J Apolinário-Hagen, T Bergmo, D Wakefield; comments to author 25.07.16; revised version received 02.11.16; accepted 25.11.16; published 22.12.16.

Please cite as:
Seth P, Abu-Abed MI, Kapoor V, Nicholson K, Agarwal G
Email Between Patient and Provider: Assessing the Attitudes and Perspectives of 624 Primary Health Care Patients
JMIR Med Inform 2016;4(4):e42
URL: http://medinform.jmir.org/2016/4/e42/
doi:10.2196/medinform.5853
PMID:28007688

©Puneet Seth, Mohamed Ismail Abu-Abed, Vikram Kapoor, Kathryn Nicholson, Gina Agarwal. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 22.12.2016. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org