JMIR Medical Informatics

Impact Factor (2023): 3.1
Volume 4 (2016), Issue 3 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Original Papers

Original Paper

Forecasting Daily Patient Outflow From a Ward Having No Real-Time Clinical Data

Shivapratap Gopakumar[1], MTech; Truyen Tran[1], PhD; Wei Luo[1], PhD; Dinh Phung[1], PhD; Svetha Venkatesh[1], PhD

Centre for Pattern Recognition and Data Analytics, Deakin University, Geelong Waurn Ponds, Australia

Corresponding Author:
Shivapratap Gopakumar, MTech
Centre for Pattern Recognition and Data Analytics
Deakin University
Building KA
75 Pigdons Road
Geelong Waurn Ponds, 3216
Australia
Phone: 61 3 5227 1266
Fax: 61 3 5227 2028
Email: shivapratap@gmail.com

Abstract

Background: Modeling patient flow is crucial in understanding resource demand and prioritization. We study patient outflow from an open ward in an Australian hospital, where currently bed allocation is carried out by a manager relying on past experiences and looking at demand. Automatic methods that provide a reasonable estimate of total next-day discharges can aid in efficient bed management. The challenges in building such methods lie in dealing with large amounts of discharge noise introduced by the nonlinear nature of hospital procedures, and the nonavailability of real-time clinical information in wards.

Objective: Our study investigates different models to forecast the total number of next-day discharges from an open ward having no real-time clinical data.

Methods: We compared 5 popular regression algorithms to model total next-day discharges: (1) autoregressive integrated moving average (ARIMA), (2) the autoregressive moving average with exogenous variables (ARMAX), (3) k-nearest neighbor regression, (4) random forest regression, and (5) support vector regression. Although the autoregressive integrated moving average model relied on past 3-month discharges, nearest neighbor forecasting used median of similar discharges in the past in estimating next-day discharge. In addition, the ARMAX model used the day of the week and number of patients currently in ward as exogenous variables. For the random forest and support vector regression models, we designed a predictor set of 20 patient features and 88 ward-level features.

Results: Our data consisted of 12,141 patient visits over 1826 days. Forecasting quality was measured using mean forecast error, mean absolute error, symmetric mean absolute percentage error, and root mean square error. When compared with a moving average prediction model, all 5 models demonstrated superior performance with the random forests achieving 22.7% improvement in mean absolute error, for all days in the year 2014.

Conclusions: In the absence of clinical information, our study recommends using patient-level and ward-level data in predicting next-day discharges. Random forest and support vector regression models are able to use all available features from such data, resulting in superior performance over traditional autoregressive methods. An intelligent estimate of available beds in wards plays a crucial role in relieving access block in emergency departments.

(JMIR Med Inform 2016;4(3):e25) doi:10.2196/medinform.5650

KEYWORDS

patient flow; discharge planning; predictive models

Introduction

Demand for health care services has become unsustainable [1,2]. This is largely due to increase in population and life expectancy, escalating costs, increased patient expectations, and workforce issues [3]. Despite increased demands, the number of inpatient beds in hospitals has come down by 2% since the last decade [2,4]. Efficient bed management is key to meeting this rising demand and reducing health care costs.

Daily discharge rate can be a potential real-time indicator of operational efficiency [5]. From a ward-level perspective, a good estimate of next-day discharges will enable hospital staff to foresee potential problems such as changes in number of available beds and changes in number of required staff. Efficient forecasting reduces bed crisis and improves resource allocation. This foresight can help accelerate discharge preparation, which has huge cost on clinical staff and educating patients and family, requiring postdischarge planning [6,7]. However, studying patient flow from general wards offers several challenges.

Ward-level discharges incorporate far greater hospital dynamics that are often nonlinear [8]. Accessing real-time clinical information in wards can be difficult because of administrative and procedural barriers, such data may not be available for predictive applications. Because the diagnosis coding is performed after discharge, there is little information about medical condition or variation in care quality in real time. In addition, factors other than patient condition play a role in discharge decisions [5,9,10].

The current practice of bed allocation in general wards of most hospitals involves a hospital staff/team, who use past information and experience, to schedule and assign beds [11]. Modern machine learning techniques can be used to aid such decisions and help understand the underlying process. As an example, Figure 1 illustrates a decision tree trained on past discharges and ward occupancy statistics, which models the daily discharge pattern from an open ward in a regional Australian hospital. Although the absence of patient medical information affected forecast performance, the decision rules provide important insight into the discharge process.

Motivated by this result, we address the open problem of forecasting daily discharges from a ward with no real-time clinical data. Specifically, we compare the forecasting performance of 5 popular regression models: (1) the classical autoregressive integrated moving average (ARIMA), (2) the autoregressive moving average with exogenous variables (ARMAX), (3) k-nearest neighbor (kNN) regression, (4) random forest (RF) regression, and (v) support vector regression (SVR). Our experiments were conducted on commonly available data from a recovery ward (heath wing 5) in Barwon Health, a regional hospital in Victoria, Australia. The ARIMA and kNN models are built from daily discharges from ward. To account for the seasonal nature of discharges, the ARMAX model included day of the week and ward occupancy statistics. We identified and constructed 20 ward-level and 88 patient-level predictors to derive the RF and SVR models.

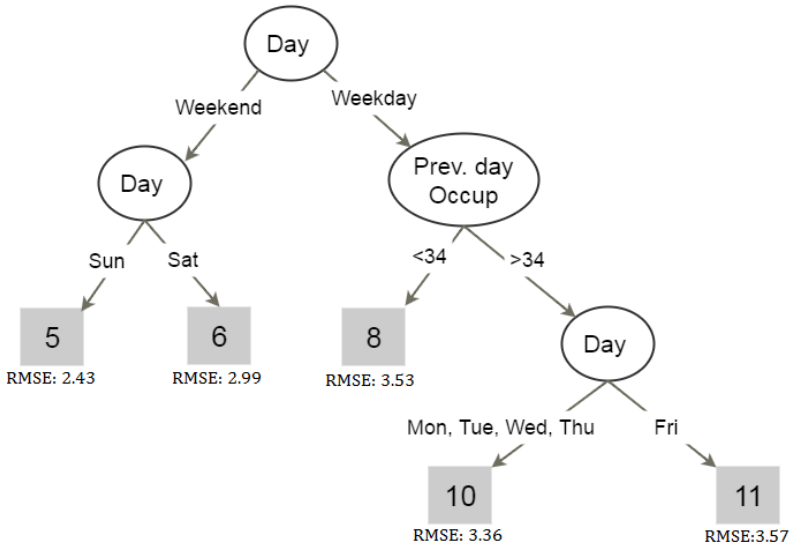
Forecasting accuracy was measured using 3 metrics on a held out set of 2511 patient visits in the year 2014. When compared with a naive forecasting method of using the mean of last week discharges, we demonstrate through our experiments that (1) using regression methods for forecasting discharge outperforms naive forecasting, (2) SVR and RF models outperform the autoregressive methods and kNN, (3) an RF model derived from 108 features has the minimum error for next-day forecasts.

The significance of our study is in identifying the importance of foreseeing available beds in wards, which could help relieve emergency access block [12].

Patient length of stay directly contributes to hospital costs and resource allocation. Long-term forecasting in health care aims to model bed and staffing needs over a period of months to years. Cote and Tucker categorize the common methods in health care demand forecasting as percent adjustment, 12-month moving average, trendline, and seasonalized forecast [13]. Although each of these methods is built from historical demand, seasonalized forecasting provides more realistic results as it takes into account the seasonal variations and trends in the data. Mackay and Lee [3] advise modeling the patient flow in health care institutions for tactical and strategic forecasting. To this end, compartmental modeling [14,15], queuing models [16,17] and simulation models [17-20] have been applied to analyze patient flow. To understand long-term patient flow, studies analyze metrics such as bed occupancy [3,8,14,19,21,22], patient arrivals [23], and individual patient length of stay [19,24-27].

On the other hand, our work implements short-term forecasting. The short-term forecasting methods are concerned with hourly and daily forecasts from a single unit in a care environment. The most popular unit of interest is the emergency or acute care department because this is often a key performance indicator metric in assessing quality of care [28,29].

Figure 1. Decision tree modeling of total discharges from an open ward from day of the week and ward occupancy (previous day occupation) data for 5 years. The leaves represent total number of patient discharges.



Time Series and Smoothing Methods

When looking at discharges as time series, autoregressive moving average models are the most popular [30-32]. Exponential smoothing techniques have also been used to forecast monthly [33] and daily patient flow [34].

Jones et al used the classical ARIMA to forecast daily bed occupancy in emergency department of a European hospital [30]. The model which included seasonality terms demonstrated reasonable performance to predict bed occupancy. The authors speculated whether nonlinear forecasting techniques could improve over ARIMA. A recent study confirmed the effectiveness of this forecasting technique in a US hospital setting [35]. ARIMA models were also successfully used to forecast the number of occupied beds during a SARS outbreak in a Singapore hospital [36]. A recent study used patient attendances in a pediatric emergency department to model daily demand using ARIMA [37].

Jones et al [34] compared the ARIMA mode with exponential smoothing and artificial neural networks to forecast daily patient volumes in emergency department. The study revealed no single model to be superior and concluded that seasonal patterns play a major role in daily demand.

Simulation Methods

Modeling using simulation is typically used to study the behavior of complex systems. An early work in investigated the effects of emergency admissions on daily bed requirements in acute care, using discrete-event stochastic simulation modeling [38]. Sinreich and Marmor [39] proposed a guide for building a simulation tool based on data from emergency departments of 5 Israeli hospitals. Their method analyzed the flow of patients clustered into 8 types along with time elements. The simulation demonstrated that patient processes are better characterized by type of the patients, rather than specific hospitals visited. Yeh and Lin used a simulation model to characterize patient flow through a hospital emergency department and reduced waiting times using a genetic algorithm [40]. A similar experiment was carried out in a geriatric department using a combination of discrete event simulation and queuing model to analyze bed occupancy [19].

Regression for Forecasting

Regression models analyze the relationship between the forecasted variable and features in the data. Linear regression that encoded monthly variations was used to forecast patient admissions over a 6-month horizon and outperformed quadratic and autoregressive models [41]. Another study used clustering and Principle Component Analysis PCA to find significant predictors from patient data to model emergency length of stay using linear regression [42]. A nonlinear approach using regression trees was proposed in forecasting patient admissions which demonstrated superior performance over a neural net framework [43].

Barnes et al used 10 predictors to model real-time inpatient length of stay in a 36-bed unit using an RF model [24].

Nonlinear regression is better suited to model the changing dynamics of patient flow. To characterize the outflow of patients from the ward, we resort to regression using RF, kNN, and SVR. In the area of pattern recognition, kNNs [44] are the most effective method that exploits repeated patterns. The kNN algorithm has been successfully applied to forecast to histogram time series in financial data [45]. The nonparametric regression using kNN has been successfully demonstrated for short-term traffic forecasting [46,47] and electricity load forecasting [48,49]. However, kNN regression has not been studied for patient flow.

Another powerful and popular regression technique, SVR, uses kernel functions to map features into a higher dimensional space to perform linear regression. Though this technique has not seen much application in medical forecasting, support vector machines have been successful in financial market prediction, electricity forecasting, business forecasting, and reliability forecasting [50].

Apart from the standard autoregressive methods, we use kNN, RFs, and SVR in forecasting next-day discharges. Because

discharge patterns repeat over time, kNN regression can be applied to search for a matching pattern from past discharges. RFs and SVR regression are powerful modelling techniques requiring minimum tuning to effectively handle nonlinearity in the hospital processes.

Recently, RF forecasting was used to predict total patient discharges from a 36 bed unit in an urban hospital [24]. Apart from 4 demographic and 2 timing predictors, this study used 3 clinical predictors for patients: (1) reason for visit: identified by a physician and recorded using International Classification of Diseases: version 9 (ICD-9) diagnosis codes [51], (2) observation status: assigned to patients for monitoring purpose, and (3) pending discharge location. Total number of discharges was estimated from aggregate of individual patient length of stay.

The absence of real-time clinical information in our data makes calculating patient length of stay impossible. Instead, we resort to modelling next-day discharges by observing previous discharge patterns and examining demographics and flow characteristics in the ward.

Methods

Data

Our study used retrospective data collected from a recovery ward in Barwon Health, a large public health provider in Victoria, Australia serving about 350,000 residents. Ethics approval was obtained from the Hospital and Research Ethics Committee at Barwon Health (number 12/83) and Deakin University. The total number of available beds depended on the number of staff assigned to the ward. On average, the ward had 36 staffed beds, but fluctuated between 20 and 80 beds with varying patient flow. The physicians in the ward had no teaching responsibilities.

Table 1. Tables in hospital database used in our data collection.

Tables	Columns
Patients	1. Patient ID
	2. Age
	3. Gender
Ward Stay	1. Admission ID
	2. Name of the ward
	3. Time (entry, exit)
	4. Bed ID
Admissions	1. Patient ID
	2. Admission ID
	3. Time (admit, discharge)
	4. Patient Class (21 categories)
	5. Admission type (7 categories)

Table 2. Cohort details.

Cohort	Stats
Total patient visits	12,141
Unique patients	10,610
Length of stay (mean, median, IQR[a])	4.26, 3, 5
Discharges per day (mean, median, IQR)	8.7, 8, 5
Admissions per day (mean, median, IQR)	8.6, 8, 5
Mean ward occupancy, IQR	30.9, 4
Gender	54.8% Female
Age (mean, median)	66, 63.23

[a]IQR, interquartile range.

The data for our study came from three tables in the hospital database, as shown in Table 1. Additional real-time data that described patient condition or disease progression were unavailable because diagnosis coding using medical codes is done after discharge. Patient flow was collected for a period of 4 years. Using the admission and discharge times for each patient, we calculated the daily discharges from our ward in study. A total of 12,141 patients were admitted into the ward with a median discharge of 8 patients per day from January 1,

2010, to December 31, 2014. Table 2 summarizes the main characteristics of our data.

A time series decomposition of our data revealed strong seasonal variations and high nonlinearity in daily discharge patterns. There was a defined weekly pattern–discharge from ward peaked on Fridays and dropped significantly on weekends (see Figure 2). This seasonal nature is in tune with previous studies [9,32]. Aggregating the daily discharges into a monthly time series revealed defined monthly patterns (see Figure 3). The data displayed no significant trend. In addition, the daily discharge pattern was found to be highly nonlinear. Our forecasting methods must be able to handle such data dynamics.

We describe the following diverse methods that are applicable to forecasting under complex data dynamics: (1) ARIMA, (2) autoregressive moving, (3) forecasting using kNN discharge patterns, (4) RF, and (5) SVR. Autoregressive methods model the temporal linear correlation between nearby data points in the time series. Nearest patterns lift this linearity assumption and assumes that short periods form repeated patterns. Finally, RF and SVR look for a nonlinear functional relationship between the future outcomes and descriptors in the past.

Figure 2. Mean admissions and discharges per day from ward.

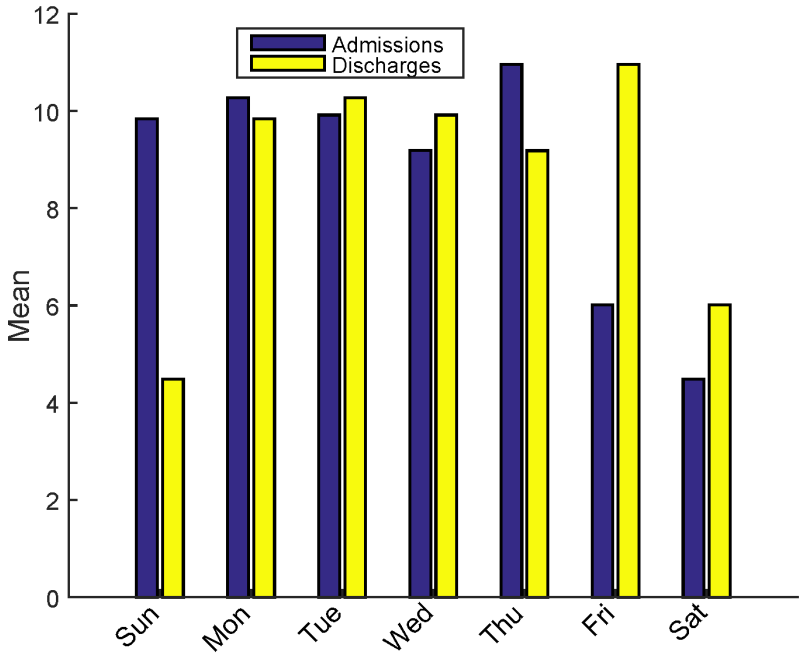
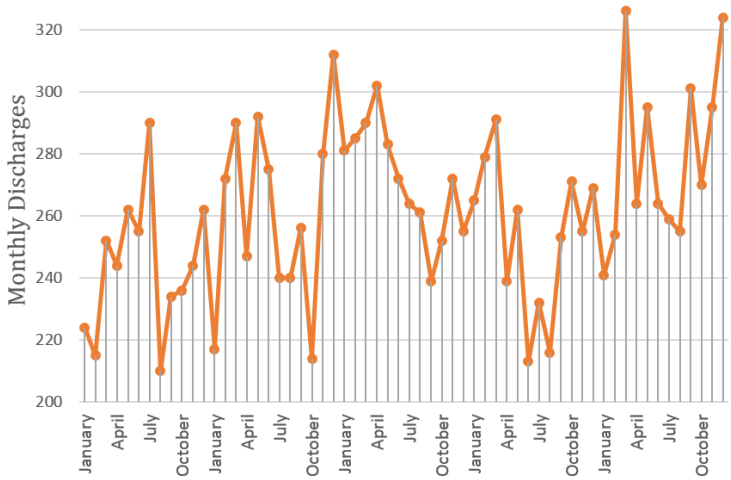


Figure 3. Time series of monthly discharges from ward.



Forecasting Methods

Autoregressive Integrated Moving Average

Time-series forecasting methods can analyze the pattern of past discharges and formulate a forecasting model from underlying temporal relationships [52]. Such models can then be used to extrapolate the discharge time series into the future. ARIMA models are widely used in time-series forecasting. Their popularity can be attributed to ease of model formulation and interpretability [53]. ARIMA models look for linear relationships in the discharge sequence to detect local trends and seasonality. However, such relationships can change over

time. ARIMA models are able to capture these changes and update themselves accordingly. This is done by combining autoregressive (AR) and moving average (MA) models. Autoregressive models formulate discharge at time t=y_t, as a linear combination of previous discharges. On the other hand, moving averages models characterize as linear combination of previous forecast errors. For ARIMA model, the discharge time series is made stationary using differencing. Let ∅ be autoregressive parameters, θ be moving average parameters, and be the forecast errors. Such an ARIMA model can be defined as shown in Figure 4, where μ is a constant. By varying p and q, we can generate different models to fit the data. Box Jenkins method [54] provides a well-defined approach for model identification and parameter estimation. In our work, we choose the auto.arima() function from the forecast package [55] in R [56] to automatically select the best model.

Figure 4. Classical ARIMA model.

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t - \sum_{i=1}^{q} \theta_q \epsilon_{t-i}$$

Autoregressive Moving Average With Exogenous Variables (ARMAX)

Dynamic regression techniques allow adding additional explanatory variables, like day of the week and number of current patients in the ward, to autoregressive models. The autoregressive moving ARMAX modifies ARIMA model by including depending external variable x_t at time t, as shown in Figure 5. We model x_t using features from the hospital database.

Figure 5. ARIMA model with exogenous variable xt.

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t - \sum_{i=1}^{q} \theta_q \epsilon_{t-i} + \beta x_t$$

Detecting Discharge Patterns Using k-Nearest Neighbors

The kNN algorithm takes advantage of the locality in data space. We assume that the next-day discharge depends on the discharges happening in previous days. Using kNN principles, we can do a regression to forecast the next-day discharge. Let y_d represent number of discharges on the current day: d. To forecast the next day discharge: y_{d+1}, we look at the discharges over the past p days as: disch_vec=[y_{d-p}: y_d]. Using Euclidean distance metric, we find k closest matches to disch_vec from the training data. An estimate of next-day discharge: $\hat{y}_{d+1}$, is calculated as a measure of the next-day discharges of the k matched patterns: $(y_{match})_i i$ (1:k). Figure 6 shows an example of kNN based forecasting. Here, disch_vec in red [y_{d-7}: y_d] results in 3 matches from the training data. For simplicity, we have plotted the matched patterns alongside disch_vec, although they had occurred in the past. The next-day forecast $\hat{y}_{d+1}$ becomes a measure of $(y_{match})_i$, where $(y_{match})_i i$ (1:3) is the $(d+1)^{th}$term of each of the matched patterns [57].

One popular method of calculating $\hat{y}_{d+1}$ is by minimizing the weighted quadratic loss (Figure 7), where w_i takes values between 0 and 1, with $\sum_{i=1}^{k} w_i = 1$. However, there are 2 main drawbacks making it less desirable for our data. First, the quadratic loss is sensitive to outliers. Second, a robust estimate of $\{w_i\}$ becomes difficult.

Our data contain significant noise, causing large variations in next-day forecasts of the k matched patterns. We illustrate this problem in Figure 8. For a given day, kNN regression returns 125 matched patterns. The next-day forecasts from each k=125 patterns displayed significant variations. In such scenario, we resort to estimating $\hat{y}_{t+1}$ by minimizing the robust loss (Figure 9).

Figure 6. k-nearest neighbor forecasting example with k=3 and P=7.

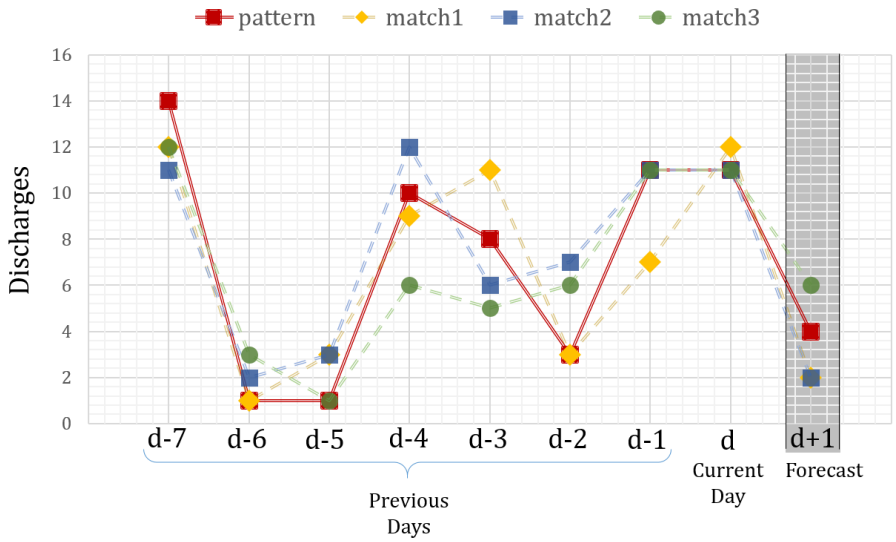


Figure 7. Calculating ŷd+1 by minimizing the weighted quadratic loss.

$$\hat{y}_{d+1} = \min_y \sum_{i=1}^{k} w_i \left((y_{\text{match}})_i - y\right)^2$$

$$= \sum_{i=1}^{k} w_i \left(y_{\text{match}}\right)_i$$

Figure 8. Scatterplot of next-day forecast using k-nearest neighbor for a given day. X-axis represents each matched nearest-neighbor pattern. Y-axis represents the next day forecast of that matched pattern.

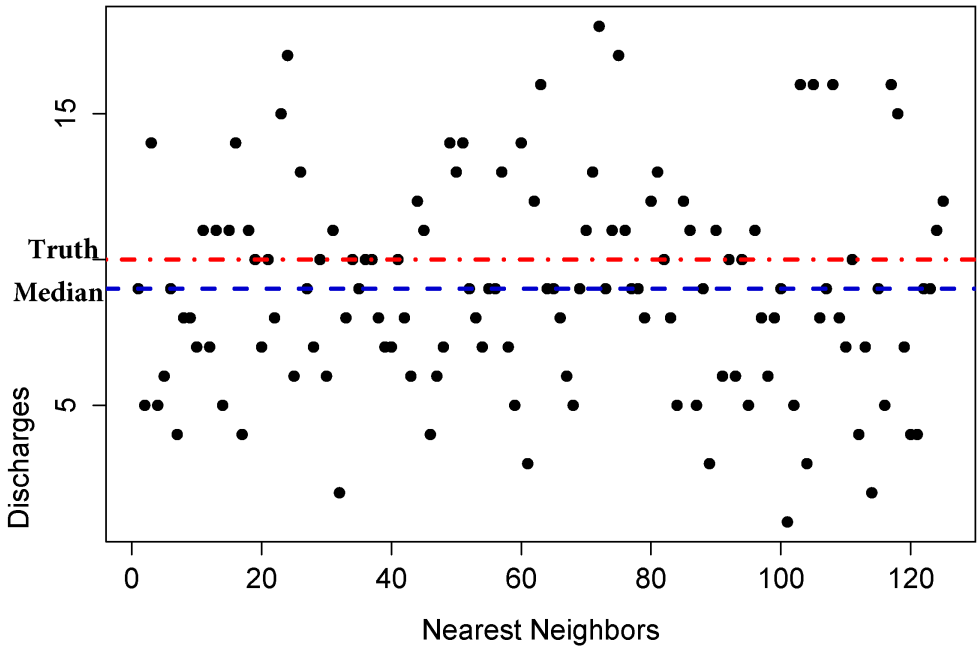


Figure 9. Estimating ŷt+1 by minimizing the robust loss.

$$\hat{y}_{t+1} = \min_y \sum_{i=1}^{k} \left|(y_{\text{match}})_i - y\right|$$

$$= \text{median}(y_{\text{match}})_{i=1 \text{ to } k}$$

Random Forest

In this approach, we assume the next-day discharge as a function of historical descriptor vector: *x*. We use each day in the past as a data point, where the next-day discharge is the outcome, and the short period before the discharge are used to derive descriptors. The RF used in this paper is currently one of the most powerful methods to model the function *y*= *f* (*x*) [58,59]. An RF is an ensemble of regression trees. A regression tree approximates a function *f* (*x*) by recursively partitioning the descriptor space. At each region R_p, the function is approximated as shown in Figure 10, where | R_p| is the number of data point falling in region R_p. The RF creates a diverse collection of random trees by varying the subsets of data points to train the trees and the subsets of descriptors at each step of space partitioning. The final outcome of RF is an average of all trees in the ensemble. Since tree growing is a highly adaptive process, it can discover any nonlinear function to any degree of approximation if given enough training data. However, the flexibility makes regression tree prone to overfitting, that is, the inability to generalize to unseen data. This requires controlling the growth by setting the number of descriptors per partitioning step, and the minimum size of region R_p.

The voting leads to great benefits: reduce the variations per tree. The randomness helps combat against overfitting. There is no assumption about the distribution of data or the form of the function (*x*). There is controllable quality of fits.

Figure 10. Random forests formulation of next day discharges (y) from historical descriptors (x).

$$f(\boldsymbol{x}) = \frac{1}{|R_p|} \sum_{x_j \in R_p} y_j$$

Support Vector Regression

The historical descriptor vector x, used in the RF model can also be used to build a SVR model [60]. Given the set of data $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$, where each $x_i \in R^m$ denotes the input descriptor for the corresponding next day forecast $y_i \in R^1$, a regression function takes the form: $\hat{y}_i = f(x_i)$. SVR works by (1) mapping the input space of x_i into a higher dimensional space using a nonlinear mapping function: φ, (2) performing a linear regression in this higher dimensional space. In general, we can express the regression function as: $f(x) = (w\varphi(x)) + b$, where, $w \in R^m$ is the weights and $b \in R^1$is the bias term. Vapnik [60] proposed the -insensitive loss function for SVR, which takes the form as shown in Equation 1 in Figure 11. The loss function L tolerates errors that are smaller than the threshold: , resulting in a "tube" around the true discharge values. Model parameters can be estimated by minimizing the cost function as shown in Equation 2 in Figure 11, where C is a constant that penalizes error in training data.

In our work, we use an RBF kernel [61] for mapping our input data to higher dimensional feature space. RBF kernels are a good choice for fitting our nonlinear discharge pattern because of its ability to map the training data to an infinite dimensional space and easy implementation. The solution to the dual formulation of SVR cost function is detailed in [60,62].

Figure 11. The SVR learning model.

$$L_\epsilon(f(\boldsymbol{x}) - \boldsymbol{y}) = \begin{cases} |f(\boldsymbol{x}) - \boldsymbol{y}|, & |f(\boldsymbol{x}) - \boldsymbol{y}| \geq \epsilon \\ 0, & \text{otherwise} \end{cases}$$

Equation 1. ϵ-insensitive loss function.

$$R = C \times \frac{1}{n} L_\epsilon(f(\boldsymbol{x}) - \boldsymbol{y}) + \frac{1}{2} \|w\|^2$$

Equation 2. SVR cost function.

Experiments

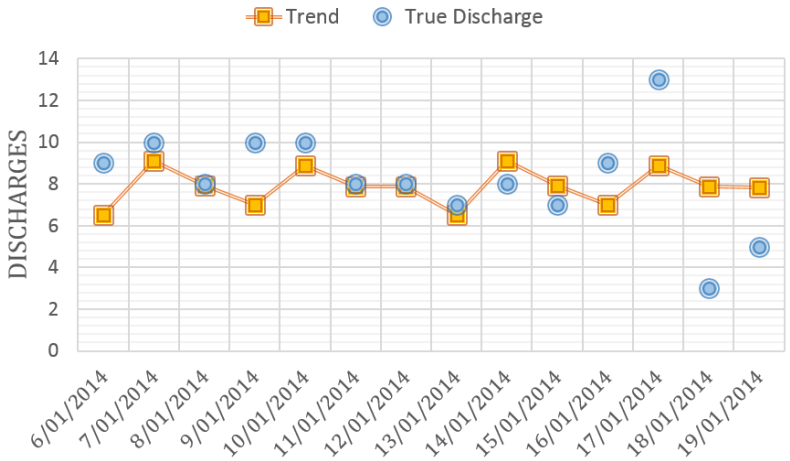
We extracted all data from the database tables (as in Table 1) for our ward in study. Patient flow was analyzed for a period of 5 years. We formatted our data as a matrix where each row corresponds to a day and each column represents a feature (descriptor). Two main groups of features were identified: (1) ward level and (2) patient level. Our feature creation process resulted in 20 ward-level and 88 patient-level predictors, as listed in Table 3. The ward-level descriptor: trend of next-day discharge was calculated by fitting a locally weighted polynomial regression [63] from past discharges. An example of this regression fitting is shown in Figure 12.

Table 3. Features constructed from ward data in hospital database.[a]

Type	Predictor	Description
Ward-level	Seasonality	Current day-of-week, current month
	Trend	Calculated using locally weighted polynomial regression from past discharges on the same weekday
	Admissions	Number of admissions during past 7 days
	Discharges	Number of discharges during past 7 days, number of discharges in previous 14th day and 21st day
	Occupancy	Ward occupancy in previous day
Patient-level	Admission type	5 categories
	Patient referral	49 categories
	Patient class	21 categories
	Age category	8 categories
	Number of wards visited	4 categories
	Elapsed length of stay	Calculated daily for each patient in the ward

[a] The random forest and support vector regression models used the full set of features. The ARMAX (autoregressive moving average with exogenous variables) model used seasonality and occupancy. All other models were derived from daily discharges.

Figure 12. An example of the discharge trend, as derived from a locally weighted polynomial regression model.



Evaluation Protocol

Our training and testing sets are separated by time. This strategy reflects the common practice of training the model using data in the past and applying it on future data. Training data consisted of 1460 days from January 1, 2010, to December 31, 2013. Testing data consisted of 365 days in the year 2014. The characteristics of the training and validation cohort are shown in Table 4. Most stays were short, around 65% of patients stayed for less than 5 days.

Table 4. Characteristics of training and validation cohorts.

Categorization	Training (2010-2013)	Testing (2014)
Total days	1460	365
Mean discharges per day	8.47	9.17
Number of admissions	9630	2511
Gender		
Male	4329 (44.9%)	1135 (45.2%)
Female	5301 (55.1%)	1376 (54.8%)
Mean age (years)	63.65	61.62
Length of stay		
1-4 days	6377 (66.22%)	1636 (65.15%)
5 or more days	3253 (33.78%)	875 (34.85%)

Baseline Forecasting

The current hospital strategy involves using past experience to foresee available beds. To compare the efficiency of our proposed approaches, we model the following baselines: (1) Naive forecasting using the last day of week discharge: since our data were found to have defined weekly patterns, we model the next day discharge as the number of discharges for the same day during previous week; (2) naive forecasting using mean of last week discharges: to better model the variation and noise in weekly discharges, we model the next-day discharge as the mean of discharges during previous 7 days; and (3) naive forecasting using mean of last 3-week discharges: to account for the monthly and weekly variations in our data, we use mean of daily discharges over the past 3 weeks to model the next-day discharge.

Measuring Forecast Performance

We compare the next-day forecasts of our proposed approaches with the baseline methods on the measures of mean forecast error, mean absolute error, symmetric mean absolute percentage error, and root mean square error [64,65]. If y_t is the measured discharge at time t, f_t is the forecasted dishcharge at time t, we can define the following:

• Mean forecast error (MFE): is used to gauge model bias and is calculated as MFE = mean(y_t- f_t)

• For an ideal model, MFE = 0. If MFE > 0, the model tends to underforecast. When MFE < 0, the model tends to overforecast.

• Mean absolute error (MAE): is the average of unsigned errors: MAE = mean$|y_t - f_t|$.

MAE indicates the absolute size of the errors.

• Root mean square error (RMSE) is a measure of the deviation of forecast errors. It is calculated as: RMSE = $\sqrt{\text{mean}(y_t - f_t)^2}$

Due to squaring and averaging, large errors tend to have more influence over RMSE. In contrast, individual errors are weighted

equally in MAE. There has been much debate on the choice of MAE or RMSE as an indicator of model performance [66,67].

•Symmetric mean absolute percentage error (sMAPE): It is scale independent and hence can be used to compare forecast performance between different data series. It overcomes 2 disadvantages of mean absolute percentage error (MAPE) namely, (1) the inability to calculate error when the true discharge is zero and (2) heavier penalties for positive errors than negative errors. sMAPE is a more robust estimate of forecast error and is calculated as: sMAPE = mean(200[$|y_t - f_t|/y_t + f_t$]). However, sMAPE ranges from −200% to 200%, giving it an ambiguous interpretation [68].

Results

Model Performance

In this section, we describe the results of comparing our different forecasting methods. The model parameters for kNN forecast, RF, and SVR models were tuned to minimize forecast errors.

For kNN regression, the optimum value of pattern length: *d* and number of nearest neighbours: *k*, was obtained by analyzing forecast RMSE for values *d* (1,100) and *k* (5,1000). Minimum RMSE of 3.77 was obtained at *d*=70 and *k*=125.

The SVR parameters *C* (penalty cost) and (amount of allowed error) were determined by choosing the best value from a grid search, that minimized the model RMSE. Similarly, the optimum number of variables in building each node of the RF was chosen by examining its effect on minimizing the out-of-bag estimate.

We compared the naive forecasting methods with our proposed 5 models using MFE, MAE, RMSE, and sMAPE. The results are summarized in Table 5, whereas Figure 13 compares the distribution of actual discharges with different model forecasts.

Table 5. Forecast accuracy of different models.

Model	Mean forecast error	Mean absolute error	Symmetric mean absolute percentage error	Root mean square error	Mean absolute error improve over naïve
Naive forecast					
Using discharge from last weekday	0.03	3.81	45.70 %	4.95	
Using mean of last week discharges	0.02	3.57	41.68 %	4.42	
Using mean of last 3-week discharges	0.04	3.44	40.14%	4.34	
ARIMA[a]	0.06	3.27	38.32 %	4.15	4.9 %
ARMAX[b]	-0.01	2.99	34.86 %	3.84	13.1 %
k-nearest neighbor	1.09	2.88	34.92 %	3.77	16.3 %
Support vector regression	0.73	2.75	32.88%	3.64	20.1 %
Random forest	0.44	2.66	31.86 %	3.49	22.7 %

[a] ARIMA: autoregressive integrated moving average

[b] ARMAX: autoregressive moving average with exogenous variables

The naive forecasts are unable to capture all variations in the data and resulted in the maximum error when compared with other models.

The variations in seasonality and trend are better captured in ARIMA and ARMAX models. The time series consisting of past 3-month discharges were used to generate the next-day discharge forecast. The ARMAX model also included the day of week and ward occupancy as exogenous variables, which resulted in better forecast performance over ARIMA.

Interestingly, kNN was more successful than ARIMA and ARMAX in capturing the variations in discharge, demonstrating about 3% improvement in MAE, when compared with ARMAX. However, the kNN model tends to under forecast (MFE = 1.09), possibly because of resorting to median values for forecast. In comparison, RF and SVR forecast models demonstrated better performance. This can be expected because they are derived from all the 108 features. However, RF demonstrated a relative improvement of 3.3 % in MAE over SVR model (see Table 5). When looking at forecast errors for each day of week, RF model confirmed better performance, as shown in Figure 14.

The process of SVR with RBF kernel maps all data into a higher dimensional space. Hence, the original features responsible for forecast cannot be recovered, and the model acts as a black box. Alternatively, RF algorithm returns an estimate of importance for each variable for regression. Examining the features with high importance could give us a better understanding of the discharge process.

Figure 13. Comparison of actual and forecasted discharges from ward for each day in 2014.

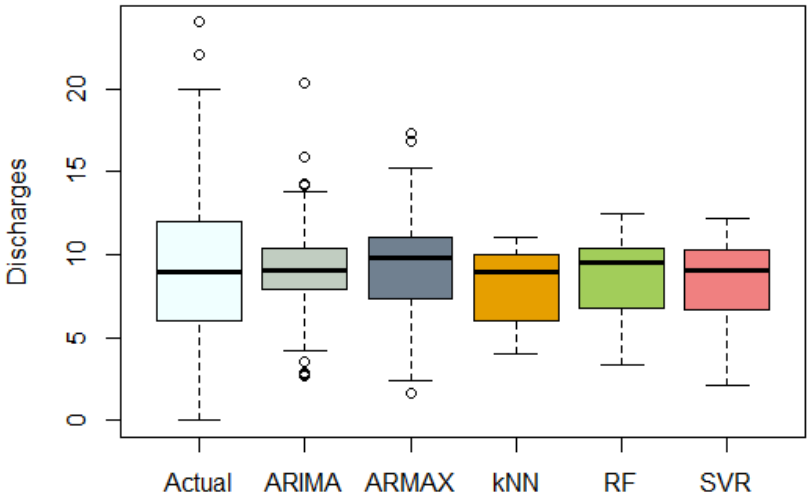
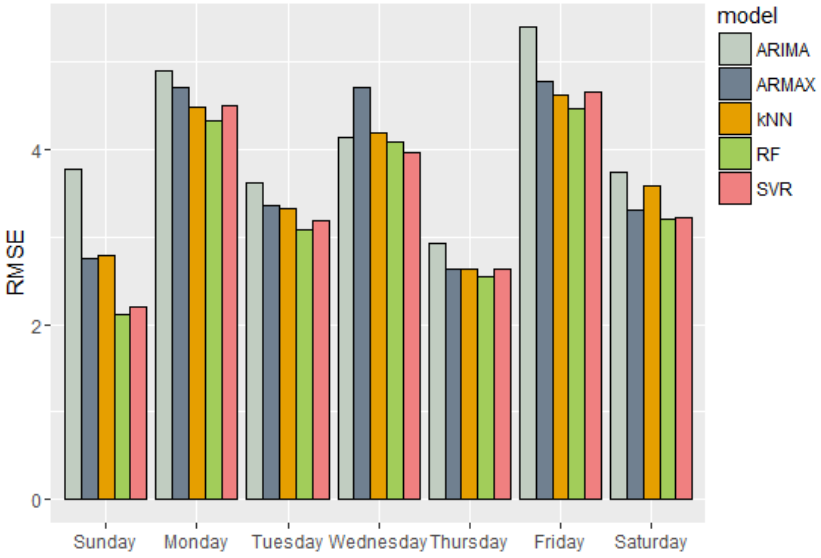


Figure 14. Forecast error in predicting each day of week in 2014.



Feature Importance in the Random Forest model

The features in random forecast model were ranked on importance scores. The top 10 significant features are described as follows. The day of week for the forecast proved to be the most important feature. Other features were number of patients in the ward during the day of forecast, the trend of discharges measured using locally weighted polynomial regression, number of discharges in past 14th day, number of discharges in past 21st day, number of patients who had visited only one previous ward, the number of males in the ward, number of patients labelled as: “public standard,” and current month of forecast.

Discussion

Principal Findings

Improved patient flow and efficient bed management is key to counter escalating service and economic pressures in hospitals. Predicting next-day discharges is crucial but has been seldom studied for general wards. When compared with emergency and acute care wards, predicting next-day discharges from a general ward is more challenging because of the nonavailability of real-time clinical information. The daily discharge pattern is seasonal and irregular. This could be attributed to management of hospital processes such as ward rounds, inpatient tests, and medication. The nonlinear nature of these processes contributes to unpredictable length of stay even in patients with similar diagnosis.

Typically, for open wards, a floor manager uses previous experience to foresee the number of available beds. In this paper, we attempt to model total number of next-day discharges using 5 methods. We have compared the forecasting performance using MAE, RMSE, and sMAPE. Our predictors are extracted from commonly available data in the hospital database. Although the kNN method is simple to implement, requiring no special expertise, software packages for other models are available for all common platforms. These models can be implemented by the analytics staff in hospital IT department and can be easily integrated into existing health information systems.

In our experiments, forecast based on RF model outperformed all other models. Forecasting error rate is 31.9% (as measured by sMAPE) which is in the same ballpark as the recent work of [24], though we had no real-time clinical information. An

RF model makes minimum assumptions about the underlying data. Hence, it is the most flexible, and at the same time, comes with great overfitting control. Similarly, SVR also demonstrated superior performance, compared with the autoregressive and kNN models. The RBF kernel maps the features into a higher dimensional space during the regression process. Hence, the physical meaning of the features is lost, making it difficult to interpret the model. Finally, RFs and SVR are able to handle more features. This extra information in the form of patient demographics and past admission and discharge statistics contributed to improve the predictive performance when compared with other models.

The kNN regression also performed well as it assumes only the locality in the data. But it is not adaptive, and thus less flexible in capturing complex patterns. The kNN regression assumes similar patterns in past discharges extrapolate to similar future discharge, which is not true for daily discharges from ward. ARMAX model outperformed the traditional ARIMA forecasts since it incorporated seasonal information as external regressors. As expected, a naive forecast of using the median of past discharges performed worst.

We noticed a weekly pattern (Figure 2) and monthly pattern (Figure 3) in discharges from the ward. Other studies have also confirmed that discharges peak on Friday and drop during weekends [5,9,10]. This "weekend effect" could be attributed to shortages in staffing or reduced availability of services like sophisticated tests and procedures [10,69]. This suggests discharges are heavily influenced by administrative reasons and staffing.

Feature importance score from an RF model helps in identifying the features contributing to the regression process. The day of forecast proved to be one of the most important features in the RF model. Other important features included trend based on nonlinear regression of past weekdays, number of discharges in the past days, ward occupancy in previous day, number of males in the ward, and number of general patients in ward.

When looking at for each day of the week, the RF and SVR model consistently outperformed other models. Sundays and Thursdays proved to be the easiest to predict for all models (Figure 14). This can be expected since these days had the least variation in our data. Fridays proved to be the most difficult to forecast. Retraining the RF model by omitting "day of the week" increased the forecast error by 1.39% (as measured by sMAPE).

Patient length of stay is inherently variable, partly due to the complex nonlinear structure of medical care [8]. The number of discharges from a ward is strongly related to the length of stay of the current patients in the ward. Hence, the variability in ward-level discharges is compounded by the variability in individual patient length of stay. In our study, the daily discharge pattern from ward shows great variation for each day of week. Apart from patient level details, we believe that a knowledge of hospital policies is also required to capture such nonlinearity.

Practical Significance

In our study, we were able to validate that the weekend patterns affect discharges from a general ward. The RF model was able to give a reasonable estimate of number of next-day discharges from the ward. Clinical staff can use this information as an aid to decisions regarding staffing and resource utilization. This foresight can also aid discharge planning such as communication and patient transfer between wards or between hospitals.

An estimate of number of free beds can also help reduce emergency department (ED) boarding time and improve patient flow [12,23]. ED boarding time is the time spent by a patient in emergency care when a bed is not available in the ward. ED boarding time severely reduces the hospital efficiency. High bed occupancy in ward directly contributes to ED overcrowding [70]. In our data, 42.81% of patients were admitted from the emergency care. An estimate of daily forecasts can be helpful in deciding the number of beds in wards to ease patient flow.

Study Limitations

We acknowledge the following limitations in our study. First, we focused only on a single ward. However, it was a ward with different patient types, and hence the results could be an indication for all general wards. Second, we did not use patient clinical data to model discharges. This was because clinical diagnosis data were available only for 42.81% of patients who came from emergency. In a general ward, clinical coding is not done in real time. However, we believe that incorporating clinical information to model patient length of stay could improve forecasting performance. Third, we did not compare our forecasts with clinicians/managing nurses. Finally, our study is retrospective. However, we have selected prediction period separated from development period. This has eliminated possible leakage and optimism.

Conclusion

This study set out to model patient outflow from an open ward with no real-time clinical information. We have demonstrated that using patient-level and ward-level features in modelling forecasts outperforms the traditional autoregressive methods. Our proposed models are built from commonly available data and hence could be easily extended to other wards. By supplementing patient-level clinical information when available, we believe that the forecasting accuracy of our models can be further improved.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments and suggestions which greatly improved the quality of the paper. This work is partially supported by the Telstra-Deakin Centre of Excellence in Big Data and Machine Learning.

Conflicts of Interest

None declared.

References

1. Kalache A, Gatti A. Active ageing: a policy framework. Adv Gerontol 2003;11:7-18. [Medline: 12820516]
2. OECD. A Disease-based Comparison of Health Systems: What is Best and at what Cost?. Paris: OECD publishing; 2003.
3. Mackay M, Lee M. Choice of models for the analysis and forecasting of hospital beds. Health Care Manag Sci 2005 Aug;8(3):221-230. [Medline: 16134435]
4. Alijani A, Hanna GB, Ziyaie D, Burns SL, Campbell KL, McMurdo ME, et al. Instrument for objective assessment of appropriateness of surgical bed occupancy: validation study. BMJ 2003 Jun 7;326(7401):1243-1244 [FREE Full text] [doi: 10.1136/bmj.326.7401.1243] [Medline: 12791738]
5. Wong H, Wu RC, Caesar M, Abrams H, Morra D. Real-time operational feedback: daily discharge rate as a novel hospital efficiency metric. Qual Saf Health Care 2010 Dec;19(6):e32. [doi: 10.1136/qshc.2010.040832] [Medline: 20724394]
6. Connolly M, Deaton C, Dodd M, Grimshaw J, Hulme T, Everitt S, et al. Discharge preparation: do healthcare professionals differ in their opinions? J Interprof Care 2010 Nov;24(6):633-643. [doi: 10.3109/13561820903418614] [Medline: 20919958]
7. Connolly M, Grimshaw J, Dodd M, Cawthorne J, Hulme T, Everitt S, et al. Systems and people under pressure: the discharge process in an acute hospital. J Clin Nurs 2009 Feb;18(4):549-558. [doi: 10.1111/j.1365-2702.2008.02551.x] [Medline: 19192004]
8. Harper P, Shahani AK. Modelling for the Planning and Management of Bed Capacities in Hospitals. The Journal of the Operational Research Society 2002;53(1):11-18 [FREE Full text]
9. Wong H, Wu RC, Tomlinson G, Caesar M, Abrams H, Carter MW, et al. How much do operational processes affect hospital inpatient discharge rates? J Public Health (Oxf) 2009 Dec;31(4):546-553 [FREE Full text] [doi: 10.1093/pubmed/fdp044] [Medline: 19465455]
10. van WC, Bell CM. Risk of death or readmission among people discharged from hospital on Fridays. CMAJ 2002 Jun 25;166(13):1672-1673 [FREE Full text] [Medline: 12126321]
11. Daniels MJ, Kuhl ME, Hager E. Forecasting Hospital Bed Availability Using Simulation and Neural Networks. In: Proceedings of IIE Annual Conference. 2005 Presented at: IIE Annual Conference and Exposition; May 14-18, 2005; Atlanta.
12. Luo W, Cao J, Gallagher M, Wiles J. Estimating the intensity of ward admission and its effect on emergency department access block. Stat Med 2013 Jul 10;32(15):2681-2694. [doi: 10.1002/sim.5684] [Medline: 23172783]
13. Côté MJ, Tucker SL. Four methodologies to improve healthcare demand forecasting. Healthc Financ Manage 2001 May;55(5):54-58. [Medline: 11351811]
14. McClean S, Millard PH. A decision support system for bed-occupancy management and planning hospitals. IMA J Math Appl Med Biol 1995;12(3-4):249-257. [Medline: 8919561]
15. McClean S, Millard PH. A three compartment model of the patient flows in a geriatric department: a decision support approach. Health Care Manag Sci 1998 Oct;1(2):159-163. [Medline: 10916595]
16. el-Darzi E, Vasilakis C, Chaussalet T, Millard PH. A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. Health Care Manag Sci 1998 Oct;1(2):143-149. [Medline: 10916593]
17. Mills TM. A mathematician goes to hospital. Australian Mathematical Society Gazette 2004;31(5):320-327.
18. Costa A, Ridley SA, Shahani AK, Harper PR, De SV, Nielsen MS. Mathematical modelling and simulation for planning critical care capacity. Anaesthesia 2003 Apr;58(4):320-327 [FREE Full text] [Medline: 12648112]
19. el-Darzi E, Vasilakis C, Chaussalet T, Millard PH. A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. Health Care Manag Sci 1998 Oct;1(2):143-149. [Medline: 10916593]
20. Hoot N, LeBlanc LJ, Jones I, Levin SR, Zhou C, Gadd CS, et al. Forecasting emergency department crowding: a discrete event simulation. Ann Emerg Med 2008 Aug;52(2):116-125. [doi: 10.1016/j.annemergmed.2007.12.011] [Medline: 18387699]
21. Mackay M. Practical experience with bed occupancy management and planning systems: an Australian view. Health Care Manag Sci 2001 Feb;4(1):47-56. [Medline: 11315885]
22. Gorunescu F, McClean SI, Millard PH. Using a queueing model to help plan bed allocation in a department of geriatric medicine. Health Care Manag Sci 2002 Nov;5(4):307-312. [Medline: 12437280]
23. Peck JS, Benneyan JC, Nightingale DJ, Gaehde SA. Predicting emergency department inpatient admissions to improve same-day patient flow. Acad Emerg Med 2012 Sep;19(9):E1045-E1054 [FREE Full text] [doi: 10.1111/j.1553-2712.2012.01435.x] [Medline: 22978731]
24. Barnes S, Hamrock E, Toerper M, Siddiqui S, Levin S. Real-time prediction of inpatient length of stay for discharge prioritization. J Am Med Inform Assoc 2016 Apr;23(e1):e2-e10. [doi: 10.1093/jamia/ocv106] [Medline: 26253131]

25. Levin SR, Harley ET, Fackler JC, Lehmann CU, Custer JW, France D, et al. Real-time forecasting of pediatric intensive care unit length of stay using computerized provider orders. Crit Care Med 2012 Nov;40(11):3058-3064. [doi: 10.1097/CCM.0b013e31825bc399] [Medline: 22824935]
26. Clark DE, Ryan LM. Concurrent prediction of hospital mortality and length of stay from risk factors on admission. Health Serv Res 2002 Jun;37(3):631-645 [FREE Full text] [Medline: 12132598]
27. Marshall A, Vasilakis C, El-Darzi E. Length of stay-based patient flow models: recent developments and future directions. Health Care Manag Sci 2005 Aug;8(3):213-220. [Medline: 16134434]
28. Kulinskaya E, Kornbrot D, Gao H. Length of stay as a performance indicator: robust statistical methodology. IMA Journal of Management Mathematics 2005;16(4):369-381.
29. Lindsay P, Schull M, Bronskill S, Anderson G. The development of indicators to measure the quality of clinical care in emergency departments following a modified-delphi approach. Acad Emerg Med 2002 Nov;9(11):1131-1139 [FREE Full text] [Medline: 12414461]
30. Jones SA, Joy MP, Pearson J. Forecasting demand of emergency care. Health Care Manag Sci 2002 Nov;5(4):297-305. [Medline: 12437279]
31. Littig SJ, Isken MW. Short term hospital occupancy prediction. Health Care Manag Sci 2007 Feb;10(1):47-66. [Medline: 17323654]
32. Lin RC, Pasupathy KS, Sir MY. Estimating Admissions and Discharges for Planning Purposes-Case of an Academic Health System. Advances in Business and Management Forecasting 2011;8:115-128.
33. Lin WT. Modeling and forecasting hospital patient movements: Univariate and multiple time series approaches. International Journal of Forecasting 1989 Jan;5(2):195-208. [doi: 10.1016/0169-2070(89)90087-3]
34. Jones SS, Thomas A, Evans RS, Welch SJ, Haug PJ, Snow GL. Forecasting daily patient volumes in the emergency department. Acad Emerg Med 2008 Feb;15(2):159-170 [FREE Full text] [doi: 10.1111/j.1553-2712.2007.00032.x] [Medline: 18275446]
35. Schweigler LM, Desmond JS, McCarthy ML, Bukowski KJ, Ionides EL, Younger JG. Forecasting models of emergency department crowding. Acad Emerg Med 2009 Apr;16(4):301-308 [FREE Full text] [doi: 10.1111/j.1553-2712.2009.00356.x] [Medline: 19210488]
36. Earnest A, Chen MI, Ng D, Sin LY. Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. BMC Health Serv Res 2005;5:36 [FREE Full text] [doi: 10.1186/1472-6963-5-36] [Medline: 15885149]
37. Kadri F, Harrou F, Chaabane S, Tahon C. Time series modelling and forecasting of emergency department overcrowding. J Med Syst 2014 Sep;38(9):107. [doi: 10.1007/s10916-014-0107-0] [Medline: 25053208]
38. Bagust A, Place M, Posnett JW. Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. BMJ 1999 Jul 17;319(7203):155-158 [FREE Full text] [Medline: 10406748]
39. Sinreich D, Marmor Y. Emergency department operations: the basis for developing a simulation tool. IIE transactions 2005;37(3):233-245.
40. Yeh JY, Lin WS. Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department. Expert Systems with Applications 2007;32(4):1073-1083.
41. Boyle J, Wallis M, Jessup M, Crilly J, Lind J, Miller P, et al. Regression forecasting of patient admission data. Conf Proc IEEE Eng Med Biol Soc 2008;2008:3819-3822. [doi: 10.1109/IEMBS.2008.4650041] [Medline: 19163544]
42. Combes C, Kadri F, Chaabane S. Predicting Hospital Length Of Stay Using Regression Models: Application To Emergency Department. 2014 Presented at: 10ème Conférence Francophone de Modélisation, Optimisation et Simulation-MOSIM; November 5-7, 2014; France p. 14.
43. Garcia KA, Chan PK. Estimating Hospital Admissions with a Randomized Regression Approach. 2012 Presented at: 11th International Conference on Machine Learning and Applications ICMLA 2012; December 12-15, 2012; Boca Raton, FL p. 179-184.
44. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory 1967;13(1):21-27.
45. Arroyo J, Mat'e C. Forecasting histogram time series with k-nearest neighbours methods. International Journal of Forecasting 2009;25(1):192-207.
46. Davis GA, Nihan NL. Nonparametric Regression and Short-Term Freeway Traffic Forecasting. Journal of Transportation Engineering 1991;117:178-188.
47. Zhang L, Liu Q, Yang W, Wei N, Dong D. An improved k-nearest neighbor model for short-term traffic flow prediction. Procedia-Social and Behavioral Sciences 2013;96:653-662.
48. Al-Qahtani FH, Crone SF. Multivariate k-nearest neighbour regression for time series data—A novel algorithm for forecasting UK electricity demand. 2013 Presented at: Neural Networks (IJCNN), The 2013 International Joint Conference; August 4-9, 2013; Dallas, TX. [doi: 10.1109/IJCNN.2013.6706742]
49. Tsakoumis AC, Vladov SS, Mladenov VM. Daily load forecasting based on previous day load. 2002 Presented at: Neural Network Applications in Electrical Engineering, 2002 (NEUREL'02); May 12-17, 2002; Honolulu p. 83-86. [doi: 10.1109/NEUREL.2002.1057973]

50. Sapankevych N, Sankar R. Time series prediction using support vector machines: a survey. IEEE Computational Intelligence Magazine 2009;4(2):24-38.
51. Centers for Disease Control and Prevention. 1978. International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) URL: http://www.cdc.gov/nchs/icd/icd9cm.htm [accessed 2016-07-12] [WebCite Cache ID 6iwJ90Tnk]
52. Chatfield C. The analysis of time series: an introduction. In: The Analysis of Time Series: An Introduction, Sixth Edition. London: Chapman & Hall/CRC; 2003.
53. Kane MJ, Price N, Scotch M, Rabinowitz P. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics 2014;15:276 [FREE Full text] [doi: 10.1186/1471-2105-15-276] [Medline: 25123979]
54. Box G, Jenkins GM, Reinsel GC. Time series analysis: forecasting and control. Englewood Cliffs, NJ: Prentice Hall; 1994.
55. Hyndman R, Khandakar JY. Automatic time series forecasting: the forecast package for R. Journal of Statistical Software 2008;26(3):1-22.
56. R Development Core Team. GBIF. Vienna, Austria; 2011. R: A Language and Environment for Statistical Computing URL: http://www.R-project.org/ [accessed 2016-07-12] [WebCite Cache ID 6iwep4Khx]
57. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician 1992;46(3):175-185.
58. Breiman L. Random forests. Machine learning 2001;45(1):5-32.
59. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.
60. Vapnik V. The nature of statistical learning theory. New York: Springer; 2000.
61. Schölkopf B, Tsuda K, Vert JP. Kernel methods in computational biology. Cambridge, MA: MIT Press; 2004.
62. Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and computing 2004;14(3):199-222.
63. Cleveland W, Grosse E, Shyun W, Chambers J, Hastie T. Local regression models. In: Statistical models in S. New York: Chapman and Hall; 1992:309-376.
64. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. International Journal of Forecasting 2006 Oct;22(4):679-688. [doi: 10.1016/j.ijforecast.2006.03.001]
65. Shcherbakov MV, Brebels A, Shcherbakova NL, Tyukov AP, Janovsky TA, Kamaev V. A survey of forecast error measures. World Applied Sciences Journal 2013;24:171-176.
66. Willmott C, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research 2005;30(1):79-82.
67. Chai T, Draxler R. Root mean square error (RMSE) or mean absolute error (MAE)? Geoscientific Model Development Discussions 2014;7:1525-1534.
68. Hyndman RJ. Another look at forecast-accuracy metrics for intermittent demand. Foresight: The International Journal of Applied Forecasting 2006;4(4):43-46.
69. Lee LH, Swensen SJ, Gorman CA, Moore RR, Wood DL. Optimizing weekend availability for sophisticated tests and procedures in a large hospital. Am J Manag Care 2005 Sep;11(9):553-558 [FREE Full text] [Medline: 16159045]
70. Forster AJ, Stiell I, Wells G, Lee AJ, van WC. The effect of hospital occupancy on emergency department length of stay and patient disposition. Acad Emerg Med 2003 Feb;10(2):127-133 [FREE Full text] [Medline: 12574009]

Abbreviations

ARIMA: autoregressive intensive moving average
ARMAX: autoregressive moving average with exogenous variables
ED: emergency department
kNN: k-nearest neighbor
MAE: mean absolute error
MAPE: mean absolute percentage error
MFE: mean forecast error
RF: random forest
RMSE: root mean square error
sMAPE: symmetric mean absolute percentage error
SVR: support vector regression

Edited by G Eysenbach; submitted 14.02.16; peer-reviewed by S Barnes, S Levin; comments to author 06.04.16; revised version received 29.05.16; accepted 21.06.16; published 21.07.16.

Please cite as:
Gopakumar S, Tran T, Luo W, Phung D, Venkatesh S
Forecasting Daily Patient Outflow From a Ward Having No Real-Time Clinical Data
JMIR Med Inform 2016;4(3):e25
URL: http://medinform.jmir.org/2016/3/e25/
doi:10.2196/medinform.5650
PMID:27444059

Original Paper

A Semi-Supervised Learning Approach to Enhance Health Care Community–Based Question Answering: A Case Study in Alcoholism

Papis Wongchaisuwat[1], MS; Diego Klabjan[1], PhD; Siddhartha Reddy Jonnalagadda[2], PhD

[1]Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, United States

[2]Divison of Health and Biomedical Informatics, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

Corresponding Author:
Papis Wongchaisuwat, MS
Department of Industrial Engineering and Management Sciences
Northwestern University
2145 Sheridan Rd
Evanston, IL, 60208
United States
Phone: 1 847 491 3383
Fax: 1 847 491 8005
Email: PapisWongchaisuwat2013@u.northwestern.edu

Abstract

Background: Community-based question answering (CQA) sites play an important role in addressing health information needs. However, a significant number of posted questions remain unanswered. Automatically answering the posted questions can provide a useful source of information for Web-based health communities.

Objective: In this study, we developed an algorithm to automatically answer health-related questions based on past questions and answers (QA). We also aimed to understand information embedded within Web-based health content that are good features in identifying valid answers.

Methods: Our proposed algorithm uses information retrieval techniques to identify candidate answers from resolved QA. To rank these candidates, we implemented a semi-supervised leaning algorithm that extracts the best answer to a question. We assessed this approach on a curated corpus from Yahoo! Answers and compared against a rule-based string similarity baseline.

Results: On our dataset, the semi-supervised learning algorithm has an accuracy of 86.2%. Unified medical language system–based (health related) features used in the model enhance the algorithm's performance by proximately 8%. A reasonably high rate of accuracy is obtained given that the data are considerably noisy. Important features distinguishing a valid answer from an invalid answer include text length, number of stop words contained in a test question, a distance between the test question and other questions in the corpus, and a number of overlapping health-related terms between questions.

Conclusions: Overall, our automated QA system based on historical QA pairs is shown to be effective according to the dataset in this case study. It is developed for general use in the health care domain, which can also be applied to other CQA sites.

(JMIR Med Inform 2016;4(3):e24) doi:10.2196/medinform.5490

KEYWORDS

machine learning; natural language processing; question answering; Web-based health communities; consumer health informatics

Introduction

A study by Pew Internet Project's research reported that 87% of US adults use the Internet, and 72% of Internet users sought health information over the Internet in the past year [1]. Other studies have also analyzed the modes in which health information is shared and its impact on consumer decision making [2,3]. Although it is known that patients are seeking information that might not be obtained during the course of their regular clinical care and valuable knowledge is publicly available in the Internet, it is not trivial for users to quickly find an accurate answer to specific questions. Consequently, community-based question answering (CQA) sites such as Yahoo! Answers tend to be a potential solution to this challenge. In CQA sites, users post a question and expect the Web-based

health community to promptly provide desirable answers. Despite a high volume of users' participation, a considerable number of questions are left unanswered, and at the same time, other questions that address the same information need are answered elsewhere. This common situation drew our attention to develop an automated system for answering both unsuccessfully answered and newly posted questions.

Substantial research exists for developing systems that address physicians' information needs at the point of care. Info buttons and other decision support tools automatically select and retrieve information from knowledge sources at the point of care [4]. Social media platforms involve exchanges of health information among peers at any place and time [5]. The advantages and disadvantages of using a social network to address the information needs compared with a search engine are described in the study by Morris et al [6]. However, limited research has been done in addressing the information needs of patients through automated approaches that synthesize the information shared across Web-based health communities. CQA systems in the health care domain address this issue.

QA systems are widely studied in both open and other restricted domains. One of the common approaches is to retrieve answers based on past QA, which is also fundamental to our work. Shtok et al [7] extracted an answer from resolved QA pairs obtained from Yahoo! Answers. Specifically, a statistical model was implemented to estimate the probability that the best answer from the past posts can satisfactorily answer a newly posted question. In addition to Shtok et al, Marom et al [8] implemented a predictive model involving a decision graph to generate help desk responses from historical email dialogues between users and help desk operators. Feng et al [9] constructed a system aiming to provide accurate responses to students' discussion board questions. An important element in these QA systems is identifying the closest (the most similar) matching between a new question and other questions in a corpus. However, this is not a trivial task because both the syntactic and semantic structure of sentences should be considered to achieve an accurate matching. A syntactic tree matching approach was proposed to tackle this problem in CQA [10]. Jeon et al [11] developed a translation-based retrieval model exploiting word relationships to determine similar questions in QA archives. Various string similarity measures were also implemented to directly compute the distance between 2 different strings [12]. A topic clustering approach was introduced to find similar questions among QA pairs [13].

An important component in QA systems is re-ranking of candidates to identify the best answer. A probabilistic answer selection framework was used to estimate the probability of an answer candidate being correct [14]. Alternatively, supervised learning-based approaches including support vector machine [15,16] and logistic regression [17] are applicable to select (rank) answers. Commonly, collecting a large number of labeled data can be very expensive or even impossible in practice. Wu et al [18] developed a novel unsupervised support vector machine classifier to overcome this problem. Other studies used different classifiers with multiple features for similar problems [19-23].

Athenikos et al [24] conducted a thorough survey reviewing state of the art in biomedical question answering systems. Morris et al [25] presented a survey study about the behavior of users in question and answer systems. Luo et al [26] developed an algorithm, SimQ, to extract similar consumer health questions based on both syntactic and semantic analysis. Vector-based distance measures were used to compute similarity score among questions. Statistical syntactic parsing and standardized unified medical language system (UMLS) were implemented to construct syntactic and semantic features, respectively. However, to effectively use the information in CQAs, we need to not only retrieve similar questions but also provide and validate potential answers. SimQ was designed to retrieve similar questions from the NetWellness [27], a health information platform that has been maintained by clinician peer reviewers. Questions collected within NetWellness tend to be clean and well structured, whereas CQA websites tend to be noisy. Wong et al has also contributed to automatically answering health-related questions based on previously solved QA pairs [28]. They provide an interactive system where the input questions are precise and short as opposed to accepting CQA questions directly as input.

In comparison to these systems, our work relies on implementing semi-supervised learning with expectation–maximization (EM) approach [29]. Semi-supervised learning uses both labeled and unlabeled data for training. Given labeled and unlabeled data, EM-based semi-supervised learning first trains an initial model using just the labeled set. This model is then used to estimate the label of each element in the unlabeled set. Next, the model is retrained using both labeled and unlabeled set with the estimated labels from the previous step. The new model is used to refine the estimated labels in the unlabeled set. These steps are iteratively repeated until the algorithm converges or reaches predefined number of iterations. In addition, we used dynamic time warping (DTW) [30] along with the vector-space distance [31] to measure similarity and incorporated biomedical concepts as additional features.

In summary, our work aims to automatically answer health-related questions based on past QA. We extracted candidate questions based on similarity measure and selected possible answers by using a semi-supervised learning algorithm. Automatically retrieving answers for questions from Web-based health communities should provide the users a potential source of health information.

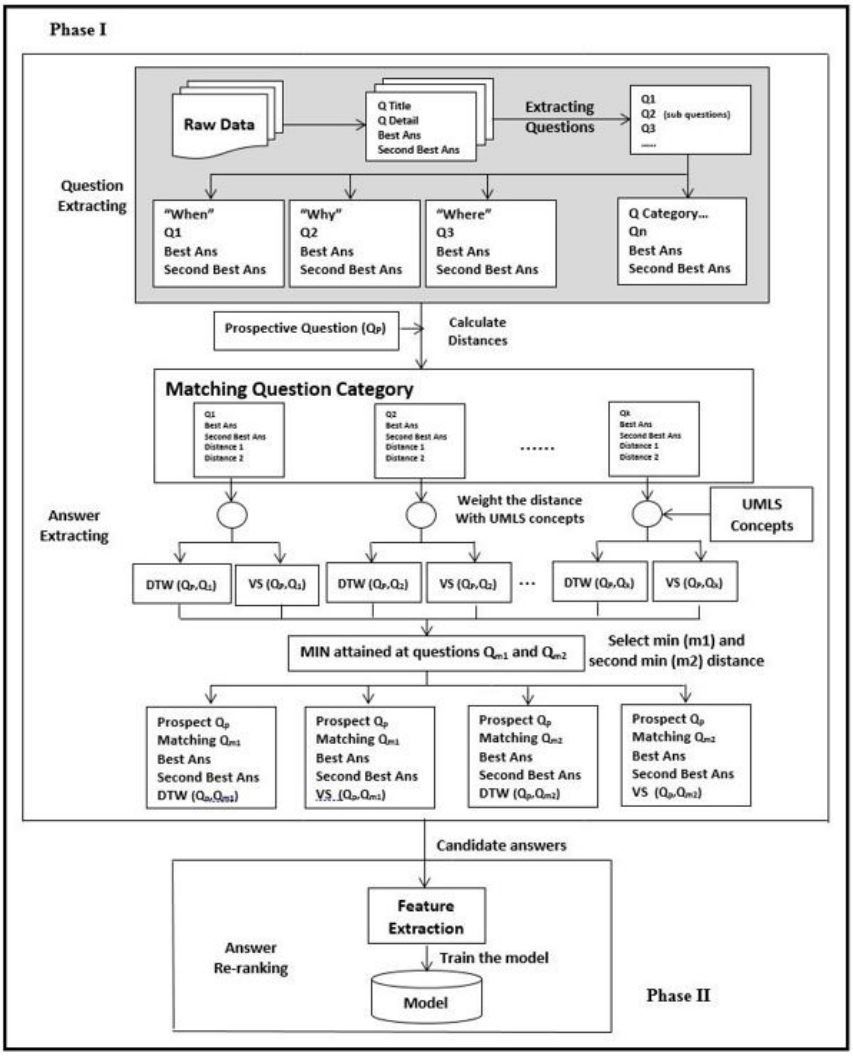
Methods

The system was built as a pipeline that involves 2 phases. The first phase implemented as a rule-based system, consists of (1) *Question Extracting*, which maps the Yahoo! Answers dataset to a data structure that includes question category, the short version of the question, and the 2 best answers; (2) *Answer Extracting*, which uses similarity measures to find answers for a question from existing QA pairs. In the second phase of *Answer Re-ranking*, we implemented supervised and semi-supervised learning models that refined the output of the first phase by screening out invalid answers and ranking the remaining valid answers.

Figure 1 depicts the system architecture and flow. In training, phase I is applied for each prospective question in the training dataset (with all other questions under a consideration corresponding to all questions in the corpus being different from the current prospective question). For test, the prospective question is a test question, and all other questions are those from the training set. In this case, phase II uses the trained model to rank the candidate answer.

We first describe the training phase. The rule-based answer extraction phase (phase I) is split into the following 2 steps:

Figure 1. Overall architecture for training the system.



Question Extracting

For this system, we assumed that each question posted on CQA sites has a question title and its description. Once users provided possible answers to the posted question, some responses were assumed to be marked as the best answer either by the question provider or community users. The second and subsequent best answers were chosen among remaining answers based on the number of likes. The raw data collected from CQA sites are unstructured and contain unnecessary text. It is essential to retrieve short and precise questions embedded in the original question title and its description (which can include up to 4-5 question sentences). Instead of using the whole question title and description that are long and verbose, we implemented a rule-based approach to capture these possible short question sentences (subquestions). These subquestions were categorized into different groups based on the words in questions. More specifically, regular expressions based on question words were used to classify subquestions, which yielded different question classes consisting of "yes-no," "what quantity," "how frequent," "when," "why," "how," "where," "who," "whose," "whom," "what," and "which," and "others." We considered subquestions, instead of full questions and descriptions, in the rest of this paper.

Answer Extracting

Given a question, it was divided into subquestions and matched with the question group using the aforementioned rule-based approach. Then, we computed the semantic distance between the prospective question and all other questions from the training sets belonging to the same group. Two distance approaches were used in our work.

1. DTW-based approach: It is based on a sequence alignment algorithm known as DTW , which uses efficient dynamic programming to calculate a distance between 2 temporal sequences. This allows us to effectively encode the word order

without adversely penalizing for missing words (such as in a relative clause). Applying it in our context, a sentence was considered as a sequence of words where the distance between each word was computed by the Levenshtein distance at a character level [32,33]. For any 2 sequences defined as $Seq_1 = < w_1^1, w_2^1, \ldots, w_m^1 >$ and $Seq_2 = < w_1^2, w_2^2, \ldots, w_n^2 >$ where m and n are the lengths of the sequences, Liu et al [30] defined the distance between 2 sequences (in our case, 2 sentences) as in the following Figure 2:

Figure 2. The distance between two sequences.

$$Seq_1 =< w_1^1, w_2^1, \ldots, w_m^1 > \text{and } Seq_2 =< w_1^2, w_2^2, \ldots, w_n^2 >$$

where $f(0,0) = 0$, $f(i, 0) = f(0, j) = \infty$, $i \in (0, m)$, $j \in (0, n)$

Here, $d(w_i^1, w_j^2)$ is the distance between 2 words computed by the Levenshtein measure.

2. Vector-space based approach: An alternative paradigm is to consider the sentences as a bag of words, represent them as points in a multidimensional space of individual words, and then calculate the distance between them. We implemented a unigram model with tf-idf weights based on the prospective question and other questions in the same category and computed the Euclidean distance measure.

We further took into account the cases that share similar medical information by multiplying the distances with a given weight parameter. The best value of the weight parameter was selected based on extensive experiments. The MetaMap tool was used to recognize UMLS concepts occurring in questions [34]. If at least 1 word in the UMLS concepts of "organic chemical" and "pharmacologic substance" occurs in both the prospective question and a training question, we reduce the distance to account for the additional semantic similarity. These UMLS concepts are specifically selected as we want to provide more weight to answers that mention a treatment approach under the intuitive assumption that most CQA users aim to seek informative advice for their illness. The set of semantic types can be expanded to capture broader concepts if different domains are considered.

The QA pairs in the training set corresponding to the smallest and the second smallest distance were extracted. Thus, we finally obtained a list of candidate answers, that is, the answers referring to smallest and second smallest questions, for each prospective question. These answers were used as the output of the baseline rule -based system. This was repeated for each question in the training set, that is, the prospective question corresponds to each question in the training set. At the end of this phase, we had triplets (Q_p, Q_t, A_t) over all questions Q_p. Note that A_t is an answer to question Q_t with $Q_t \neq Q_p$, and each Q_p yielded several such triplets.

The machine-learning phase of answer re-ranking (phase II) is described next. The goal of this phase is to rank candidate answers from the previous step and select the best answer among them. Each triple (Q_p, Q_t, A_t) is aimed to be assigned as "valid" if A_t is a valid answer to Q_p, or "invalid" otherwise. We describe how the model was trained in this section while detailed explanations (eg, number of labeled and unlabeled triplets) are provided in the section, "Results." We first selected a small random subset of triplets and labeled them manually (there were too many to label all of them in this way). Both supervised and semi-supervised learning EM models were developed to predict the answerability of newly posted question and rank candidate answers. Specifically, the semi-supervised learning model was trained on labeled and unlabeled triplets. According to the semi-supervised learning model, we first trained a supervised learning algorithm including Neural Networks with the entropy objective function (NNET), Neural Networks with the L2-norm or least squares objective function (NNET_L2), support vector machine (SVM), and logistic regression based on manually labeling outputs from the aforementioned rule-based answer extraction phase. The trained model was used to classify the unlabeled part of the outputs of phase I, and then, the classifier was retrained based on the original labeled data and a randomly selected subset of unlabeled data using the estimated labels from the previous iteration. These steps were iteratively repeated to achieve a final estimated label. The supervised approach, on the other hand, only ran a classifier on the labeled subset and finished. A 10-fold cross validation was implemented in both semi-supervised and supervised approaches. Specifically, all labeled observations were partitioned into 10 parts where 1 part was set aside as a test set. The model was fitted based on the remaining 9 parts of the labeled observations (plus the entire unlabeled part for the semi-supervised learning approach). The parameters of the semi-supervised model were obtained by using the EM algorithm previously described. The fitted model was then used to predict the responses in the part that we set aside as the test set. These steps were repeated by selecting different part to set aside as the test set. All features used in the models are illustrated based on the following example as summarized in Table 1.

Table 1. List of features used in the model.

Type of features	Features	Value
General Features	1.Text length of Q_p	5
	2.Text length of Q_t	12
	3. Number of stop words contained in Q_p	1
	4. Number of stop words contained in Q_t	5
	5. VS(Q_p, Q_t)	3.7052
	6. The difference between VS(Q_p, A_t) and VS(Q_t, A_t)	0.4303
	7. DTW(Q_p, Q_t)	29
	8. The difference between DTW(Q_p, A_t) and DTW(Q_t, A_t)	14.5
UMLS-based Features	9. Number of overlapping words in S_P and S_T	3
	10. Number of overlapping words in S_P and S_A	3
	11. Binary variable indicating whether a set of overlapping words in (S_P, S_T) and (S_P, S_A) are different	0
	12. Cardinality of the set difference of S_P and S_T	4
	13. Cardinality of the set difference of S_P and S_A	5

Example of a Triple (Q_p, Q_t, A_t)

Prospective Question

Anxiety medication for drug/alcohol addiction?

Training Question

Is chlordiazepoxide/librium a good medication for alcohol withdrawal and the associated anxiety?

Training Answer

Chlordiazepoxide has been the standard drug used for rapid alcohol detox for decades and has stood the test of time. The key word is rapid the drug should really only be given for around a week. Starting at 100 mg on day 1 and reducing the dose every day to reach zero on day 8. In my experience, it deals well with both the physical and mental symptoms of withdrawal. Looking ahead, he will still need an alternative management for his anxiety to replace the alcohol. Therapy may help, possibly in a group setting

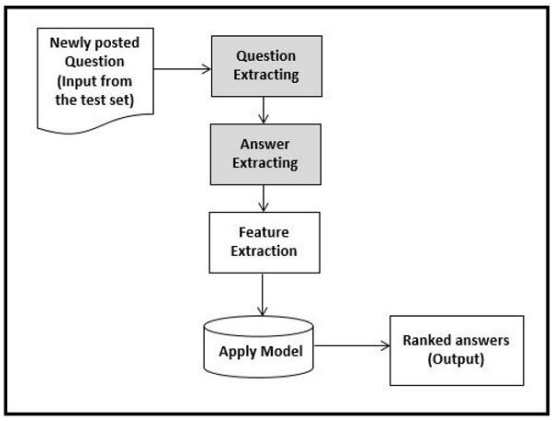
Sets S_P, S_T, and S_A are sets of terms corresponding to UMLS concepts occurred in Q_p, Q_t, and A_t, respectively. General features are taken from previous work [7], while we introduce UMLS-based features into the model. Features 9 and 10 are calculated by counting the number of words contained in both sets. To obtain features 12 and 13, we find the elements that are in only 1 of the 2 sets.

Table 2 depicts examples of annotations in the corpus. The inter-rater agreement for random instances (10% of total) assigned to 2 independent reviewers is very good (95% CI of kappa from .69 to .93). The procedure to identify an answer to a newly posted question is illustrated in Figure 3 after the usual split of the corpus in train and test.

Table 2. Corpus annotation examples.

A target question	A training question	A training answer	Label
Can fully recovered alcoholics drink again	Can a recovered alcoholic drink again?	What they say at AA is that there is no such thing as permanent recovery from alcoholism. There are alcoholics who never drink again, but never alcoholics who stop being alcoholics.	valid
Can fully recovered alcoholics drink again	If both my parents are recovered alcoholics, will I have a problem with alcohol?	Yes, there is a good chance that you could inherit a tendency towards alcoholism.	invalid
Anxiety medication for drug/alcohol addiction?	Is chlordiazepoxide/librium a good medication for alcohol withdrawal and the associated anxiety?	Chlordiazepoxide has been the standard drug used for rapid alcohol detox for decades and has stood the test of time.	valid
Anxiety medication for drug/alcohol addiction?	Negative effects of alcohol and ADHD medication?	Drinking in moderation is wise for everyone, but it is imperative for adults with ADHD.	invalid

Figure 3. Process flow of the testing step.



The following evaluation metrics are used to test the overall performance of our algorithm.

1. Question-based evaluation metrics

- For this paper, we define "overall accuracy" as ratio of the number of questions with at least 1 "correct" answer divided by total number of questions in the test set. A test question is labeled as "correct" if our algorithm predicts at least 1 valid triple correctly. For the case that there is no valid answer in the question from the gold standard, we label it as "correct" if our algorithm predicts corresponding triplets as invalid.

- The mean reciprocal rank (MRR) with test questions Q is defined as Figure 4.

where $rank_i$ is the position of a valid instance in manually sorted probabilities from the model. If there are more than 1 valid instances in any question, minimum value of $rank_i$ is used.

2. Triple-based evaluation metrics

Precision, recall, and the F1-score can be used as standard measures for binary classification. We do not measure accuracy and receiver operating characteristic curves because the dataset is heavily imbalanced.

Figure 4. The Mean Reciprocal Rank (MRR) with a set of test questions Q.

$$f(0,0) = 0, f(i,0) = f(0,j) = \infty, i \in (0,m), j \in (0,n)$$

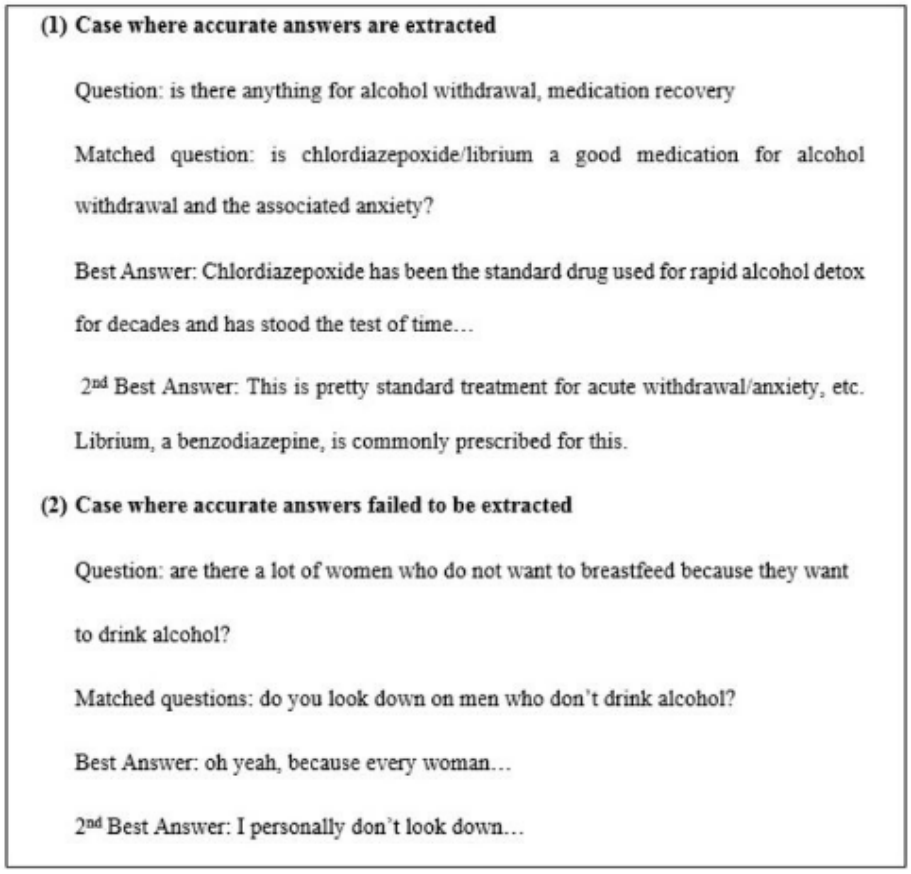
Results

To test the algorithm, we obtained a total of 4216 alcoholism-related QA threads from Yahoo! Answers. The sample outputs from our algorithm are shown in Figure 5, which indicates how our system could potentially be used by Web-based advice seekers. To extract initial candidate answers in the rule-based answer extraction, our algorithm returns 8 instances for each prospective question (obtained from 2 different similarity measures where we extract at least 2 closest questions for each measure with 2 answers for each question). An example of output reported from the rule-based answer extraction is depicted in Figure 6.

Figure 5. System output.

Input (a newly posted question)	Output (a possible answer)
Question title: Is there anything for alcohol withdrawal, medication recovery **Question description:** Just like how there is some medication for drug addicts, like Suboxone, Methadone... etc. I wonder if theres anything like that for alcoholic? ..weed? lol.. or is there? - just wondering, thanks.	chlordiazepoxide has been the standard drug used for rapid alcohol detox for decades and has stood the test of time. the key word is rapid the drug should really only be given for around a week. starting at 100 mg on day one and reducing the dose every day to reach zero on day 8. in my experience it deals well with both the physical and mental symptoms of withdrawal. looking ahead he will still need an alternative management for his anxiety to replace the alcohol. therapy may help, possibly in a group setting.
Question title: Symtons off alcohol abuse **Question description:** what side affecs can i expect when in stop drinking	i hope your meaning long term? not if you drink one night then suddenly stop all these return to normal, because they reactions will be piss poor when your pissed. fitness: will be down, the same as when you eat unhealthy etc.. you need to exercise to improve this again to a good standard. liver: if you havent abused alcohol for long & your still young, your liver can return to health naturally.. if it has been abused for years, you may have some liver disease, where it cant recoved.. reflexes: might be slightly worse of.. but shouldnt be affected too much, after normal exercise etc, should be back to normal.
Question title: Alcohol effects baby **Question description:** So I'm four months preggers and evryones been telling me tht i shud not drink becuz its bad for the baby. i know pregnant women shudnt drink but thts just becuz they might do something they might not normally do thts bad, lik sleep with a random person or drive and crash. but wat if i drink and my friend watches over me and makes sure i don't do anything bad? Or shud i just not risk it?	no! if you suspect youre pregnant, dont go anywhere near alcohol -period. youre going to ruin the life of an innocent child who deserves more if you do.

Figure 6. An example result returned from the algorithm to determine candidate answers.



A randomly selected set of 220 threads were manually annotated and used as labeled questions. Overall, 119 of 220 questions, or 54.1%, have valid answers among those extracted in the rule-based answer extraction phase. After retrieving candidate answers, we further aim to re-rank them and select the best answer (if there is a valid answer). Note that each question corresponds to several candidate answers and thus multiple triplets (Q_p, Q_t, A_t). If at least 1 triplet is labeled as "valid," the corresponding question is also labeled as "valid." Specifically, the semi-supervised learning model (EM) was trained on 1553 labeled triplets (corresponding to 220 manually labeled questions) and 10,000 unlabeled triplets. In the training data of 1553 labeled triplets, 297 triplets were manually labeled as "valid" and 1256 as "invalid." The typical 10-fold cross validation was implemented to validate the model.

We included all features listed in Table 1 in the models. To indicate a significance of each feature, we analyzed the feature set by using information gain. The information gain is based on the entropy function, which is closely related with the objective function of the neural network NNET and logistic regression classifiers. The most influential features are the number of stop words contained in Q_p, the text length, the distance of (Q_p, Q_t), and the number of overlapping UMLS words between Q_p and Q_t, that is, in S_P and S_T. All information gains for these significant features are listed in Table 3.

Table 3. Information gain score of 5 significant features

Features	Information gain
1. Number of stop words contained in Q_p	0.0912
2. Text length of Q_p	0.0804
3. DTW(Q_p, Q_t)	0.0395
4. Number of overlapping words in S_P and S_T	0.0393
5. VS(Q_p, Q_t)	0.0350

The best model was selected by varying the cutoff probability of being valid or invalid to obtain the maximum F1-score. We selected NNET, NNET_L2, SVM, and logistic regression approaches to train the model on a subset. For the SVM classifier, the probability was obtained by fitting a logistic distribution using maximum likelihood to the decision values provided by SVM.

The semi-supervised learning (EM) algorithm with 1 iteration trained with NNET_L2 gave the best performance for MRR and F1-score with a reasonable value of overall accuracy,

whereas NNET performs best for overall accuracy, as listed in Table 4. Each value in the table is the average across 100 different runs based on different random numbers in the algorithms and the test/train splits (details provided in the following section). In Table 4, the numbers in bold represent the best value among different models and classifiers for each evaluation metric. The confusion matrices for 1 iteration of EM trained with 4 different classification models are provided in Figure 7.

Figure 7. The confusion matrices for 1 iteration of EM trained with NNET, NNET_L2, SVM, and LOG.

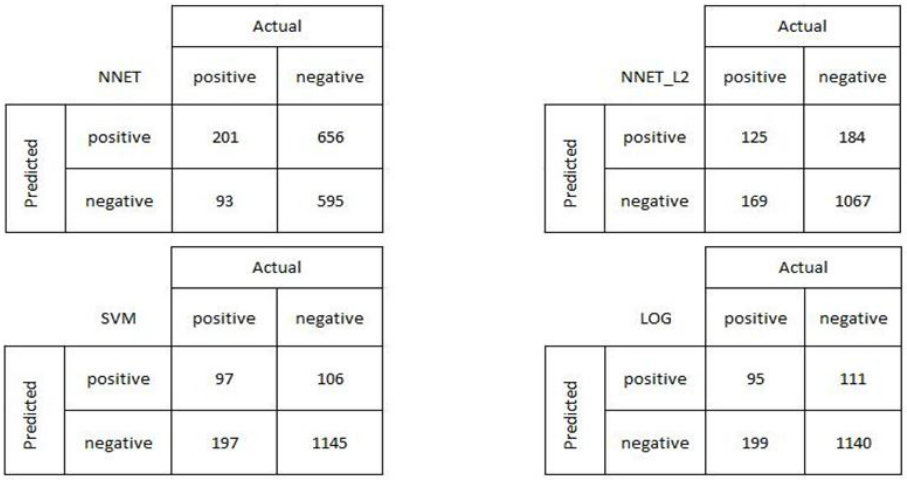


Table 4. Evaluation metrics.

Evaluation Metrics	Supervised learning				Semi-supervised learning (EM)							
	NNET	NNET_L2	SVM[a]	LOG[b]	1 iteration				10 iterations			
					NNET	NNET_L2	SVM	LOG	NNET	NNET_L2	SVM	LOG
Overall accuracy	0.5818	0.6993	0.6305	0.6245	0.8623	0.7105	0.6774	0.6473	0.8491	0.71	0.6783	0.6478
MRR[c]	0.4216	0.5534	0.6224	0.6336	0.5686	0.6339	0.631	0.6266	0.5681	0.6332	0.6313	0.628
F1-score	0.1	0.3786	0.3045	0.3214	0.3222	0.3996	0.3667	0.3622	0.316	0.3977	0.3656	0.3626
Precision	0.0746	0.3614	0.4803	0.5073	0.2294	0.3981	0.4493	0.4421	0.2219	0.3942	0.4478	0.44
Recall	0.1433	0.4	0.241	0.2659	0.6801	0.4214	0.3239	0.3224	0.6562	0.4209	0.3229	0.3233

[a]SVM: support vector machine.

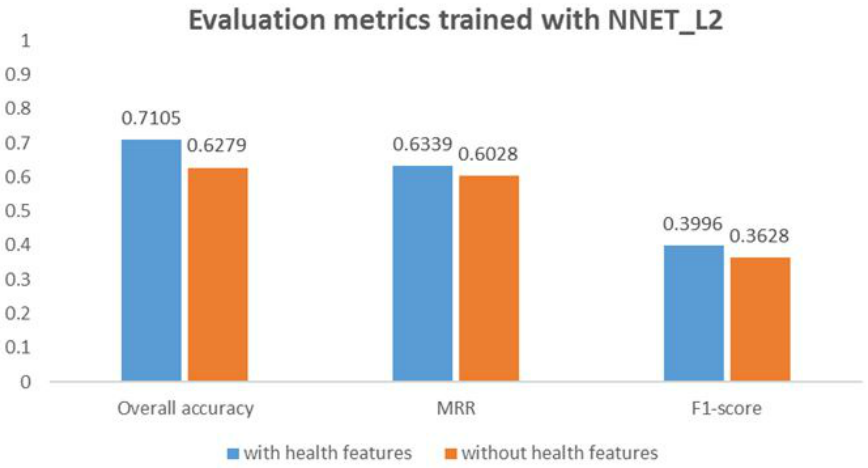
[b]LOG: logistic regression.

[c]MRR: mean reciprocal rank.

We performed 2 types of statistical hypothesis tests (*t*-tests) at the .05 level (95% CI) to determine if 2 sets of evaluation metrics among the F1-score, overall accuracy, and MRR, obtained from different settings are significantly different from each other. First, randomness occurs within an algorithm such as the randomness in the stochastic gradient approach. Second, we consider randomness of assigning the test set, that is, the training and test sets in 10-fold cross validation are randomly assigned. We performed both types of the hypothesis tests for all possible comparisons including the model implemented (pure classification vs semi-supervised), and among the 4 different classifiers based on the numbers reported in Table 4. Overall, the semi-supervised learning model is statistically significantly better than the corresponding supervised version for all evaluation metrics. This conclusion holds for both tests. Comparing between 1 and 10 EM iterations, the evaluation metrics are not statistically different from each other. This implies that the model parameters tuned by the EM algorithm are very close to the optimal values within 1 iteration.

We are also interested in understanding whether UMLS-based features (feature 9-13 listed in Table 1) play a role in predicting the validity of a candidate answer. Hence, we trained another model, which excludes all UMLS-based features, and compared the results (obtained from 1 iterations of EM trained with NNET_L2) with those from the original model as illustrated in Figure 8. The statistical tests at the .05 level showed significantly difference between the 2 models (with vs without UMLS-based features) for the 3 evaluation metrics. With UMLS-based features, the model gave a better performance, which is consistent across all evaluation metrics. This implies that these features played a role in distinguishing between valid and invalid answers.

Figure 8. Performance between the original and adjusted model to test significance of UMLS-based features (health features).



Discussion

In this paper, we developed an automated QA system by using previously resolved QA pairs from a CQA site and evaluated it. Although we used Yahoo! Answers as a data source, our algorithm can be adapted and applied to other CQA sites, in particular those related to health care where UMLS applies. Among different models and classifiers experimented, EM semi-supervised learning is better than pure supervised learning, and 1 iteration of EM generally performs better than other models. Specifically, 1 iteration of EM with NNET gives the best performance in term of accuracy. NNET_L2 with 1 iteration of EM performs best in terms of the MRR and F1-scores. The NNET_L2 with 1 EM iteration is recommended to be used based on the case study data. Overall, the best model achieves an 86.2% accuracy and a 0.4 F1-score, which are significant given that the problem is challenging and the data are imperfect. Internet users typically provide responses in an ill-formed fashion. Our data also consist of a significant number of complex questions, for example, a user discusses about his or her situation in 10 to 20 sentences and then asks whether he or she is an alcoholic. Moreover, some questions are very detailed; for example, the percentage of alcohol resulting from a given combination of chemical components. There is a trade-off between precision and recall. Some of these values listed in Table 4 are small as we aim to find a good balance between the 2 values. We intentionally maximize the F1-score, which is a representative of both values. Precision and recall are reported in Table 4 for completeness. A comparison between the rule-based approach in the first phrase and the semi-supervised learning model in the second phrase reveals a significant improvement. The semi-supervised approach improves the accuracy of the model by 30% (approximately from 55% to 86%).

Comparing with Luo et al [26] who retrieved the similar questions based on the distance measure, we relied on this idea with different approaches. To compute the similarity score between questions, we used the DTW measure instead of relying on the vector-based distance measure. Luo et al used matching questions with information in data sources that are written and reviewed by experts; we strictly use only data from Yahoo! Answers, which are very noisy. For this reason, the syntactic features proposed by Luo et al might not be useful in our model. Unfortunately, not all libraries used in Luo et al's implementation are publicly available, and thus, direct comparison of the accuracy is not possible.

Shtok et al [7] used resolved QA pairs to reduce the rate of unanswered questions in Yahoo! Answers. The experiment in Shtok et al was also tested with health-related questions, and the accuracy as measured by the F1-score was 0.32. Our method, which trained a semi-supervised learning model with a smaller amount of manually labeled data compared with a supervised learning model used in [7], resulted in 0.4 F1-score. A better performance might be because of several reasons. First, we categorized questions in a corpus into different groups based on question keywords. Instead of computing the distance between a test question and all other questions in the corpus, categorizing questions reduces the scope of questions an algorithm needs to search. As we categorize collected questions into different groups based on question keywords, latent topics and "wh" question matching features used in Shtok's study are not valuable in our context. Second, our algorithm also used multiple features related to the UMLS medical topics to enhance the model's performance when applied within the health domain where the Shtok's system was designed for a more general usage. Although Shtok et al. relied on cosine distance, the Euclidean distance performed better in our evaluation. Among distance measures used in our work, more valid answers can be correctly identified with the DTW-based approach than the vector similarity measure, which can be observed when manually annotating the output from the rule-based answer extraction. In addition, our algorithm extracted multiple candidate answers retrieved from 2 closest QA pairs for each distance metric and the 2 best answers for each question. In each QA pair, both the best and the second best answer were extracted compared with Shtok et al where only the best answer was extracted. Finally, we implemented semi-supervised learning to gain benefits from unlabeled data, whereas Shtok et al only relied on a supervised learning model in the re-ranking phase.

Using a semi-supervised learning model that leverages unlabeled data is reasonable against other traditional supervised learning

models because obtaining labeled data is very expensive and time consuming in practice. As the features of the machine-learning algorithm are not specific to alcoholism, our system should be applicable for other related topics. On the other hand, it would be possible to increase the accuracy for "alcoholism" if we use specific features such as concepts related to alcoholism.

In summary, the main novelty and advantages of our work against other works include the DTW-based distance approach, UMLS-based features, the semi-supervised learning algorithm, and the dataset used in the study. We introduce novel distance measures, the DTW-based approach that performs better than the typical vector-space distance method. UMLS-based features are included to enhance the model applied in the health care domain in addition to the general features in the study by the study by Shtok et al [7]. Our system is trained and tested only on the Web-based information without any additional sources. Further, obtaining the annotation from Web-based data can be very difficult and time consuming. This stresses the significance of using semi-supervised learning rather than a typical supervised learning algorithm.

For the machine-learning component, the distance between a test question and other questions in the training dataset is important in distinguishing valid and invalid answers. The closer the distance is, the higher the chance of the corresponding answer being valid. Matching UMLS terms, which imply a closer similarity between questions, plays a role in determining the validity of the answer. Although UMLS-based features show lower information gain, the model with these features included is significantly better across all evaluation metrics. The overall accuracy is improved by 8% when these features are included. Information gain shows that number of stop words contained in a test question and the underling text length are the best indicators for differentiating between valid and invalid answers. We note that the number of content-rich words, represented as text length minus the number of stop words, is also taken indirectly into account by these 2 features. We fitted the model without the number of stop words feature compared with the full model. Although these 2 models are not statistically different, we include the number of stop words feature in the model as previously done by Shtok et al [7].

Limitations and Future Work

The main limitation of our work is the lack of assessment of the model's generalizability. Although our algorithm is generic and does not include any features that are specific to the topic of alcoholism, we have not validated it in different domains as we do not have available data. Approximately 30% (obtained from a preliminary observation) of all questions cannot be answered based on existing answers; some of these questions also require additional resources that are more technical and reliable, such as medical textbooks, journals, and guidelines.

Conclusions

The question-answering system developed in this work achieves reasonably good performance in extracting and ranking answers to questions posted in CQA sites. Our work is a promising approach for automatically answering alcoholism-related questions obtained from CQA sites based only on past QA that is used as a case study. In addition, our system can potentially be applied to other health care domain questions asked by Web-based health care communities. The system and the gold standard corpus are available on GitHub [35].

Acknowledgments

The authors are grateful to Dr. Jina Huh from Michigan State University for providing the Yahoo! Answers dataset and Dr. Kalpana Raja from Northwestern University for helping with UMLS. This work was partly supported by funding from the National Library of Medicine grant R00LM011389.

Conflicts of Interest

None declared.

References

1. Fox S, Duggan M. Pew Research Center. 2013. Health Online 2013 URL: http://www.pewinternet.org/2013/01/15/health-online-2013/ [WebCite Cache ID 6eAaGTAoU]
2. Lau AY, Coiera E. Impact of web searching and social feedback on consumer decision making: a prospective online experiment. J Med Internet Res 2008;10(1):e2 [FREE Full text] [doi: 10.2196/jmir.963] [Medline: 18244893]
3. Nath C, Huh J, Adupa A, Jonnalagadda S. Website sharing in online health communities: A descriptive analysis. J Med Internet Res 2016;18(1):e11 [FREE Full text] [doi: 10.2196/jmir.5237] [Medline: 26764193]
4. Cimino J, Yu H, Del FG. Infobuttons and point of care access to knowledge. Clinical Decision Support 2007:345-372.
5. Burton S, Tanner K, Giraud-Carrier C. Leveraging social networks for anytime-anyplace health information. Network modeling analysis in health informatics and bioinformatics 2012;1(4):173-181.
6. Morris M, Jaime T, Panovich K. A comparison of information seeking using search engines and social networks. 2010 Jan 01 Presented at: Proceedings of 4th international AAAI conference on weblogssocial media; May 23-26, 2010; Washington, DC p. 291-294.

7. Shtok A, Dror G, Maarek Y, Szpektor I. Learning from the past: answering new questions with past answers. 2012 Apr 16 Presented at: Proceedings of the 21st international conference on world wide web; April 16-20, 2012; Lyon, France p. 759-768. [doi: 10.1145/2187836.2187939]
8. Marom Y, Zukerman I. A predictive approach to help-desk response generation. 2007 Jan 06 Presented at: 20th international joint conference on artificial intelligence; January 6-12, 2007; Hyderabad, India p. 1665-1670.
9. Feng D, Shaw E, Kim J, Hovy E. An intelligent discussion-bot for answering student queries in threaded discussions. 2006 Jan 29 Presented at: Proceedings of the 11th international conference on intelligent user interfaces; January 29 - February 01, 2006; Sydney, Australia p. 171-177.
10. Wang K, Ming Z, Chua T. A Syntactic tree matching approach to finding similar questions in community-based QA services. 2009 Jul 19 Presented at: Proceedings 32nd annual international ACM SIGIR conference on research and development in information retrieval; July 19-23, 2009; Boston, MA p. 187-194.
11. Jeon J, Croft W, Lee J. Finding similar questions in large question and answer archives. 2005 Oct 31 Presented at: Proceedings of the 14th ACM international conference on information and knowledge management; October 31 - November 05, 2005; Bremen, Germany p. 84-90.
12. Bernhard D, Gurevych I. Answering learners' questions by retrieving question paraphrases from social Q&A sites. 2008 Jun 19 Presented at: Proceedings of the third workshop on innovative use of NLP for building educational applications; June 19, 2008; Columbus, OH p. 44-52.
13. Zhang W, Liu T, Yang Y, Cao L, Zhang Y, Ji R. A topic clustering approach to finding similar questions from large question and answer archives. PLoS One 2014;9(3):e71511 [FREE Full text] [doi: 10.1371/journal.pone.0071511] [Medline: 24595052]
14. Ko J, Si L, Nyberg E. A probabilistic framework for answer selection in question answering. 2007 Apr 22 Presented at: Proceedings of human language technology conference of the North American chapter of the association of computational linguistics; April 22-27, 2007; Rochester, NY p. 524-531.
15. Moschitti A, Quarteroni S. Linguistic kernels for answer re-ranking in question answering systems. Information Processing & Management 2011;47(6):825-842. [doi: 10.1016/j.ipm.2010.06.002]
16. Suzuki J, Sasaki Y, Maeda E. SVM answer selection for open-domain question answering. 2002 Aug 26 Presented at: Proceedings of the 19th international conference on computational linguistics; August 26-30, 2002; Taipei, Taiwan p. 1-7.
17. Blooma M, Chua A, Goh D. Selection of the best answer in CQA services. 2010 Apr 12 Presented at: Proceedings of the 7th international conference on information technology; April 12-14, 2010; Las Vegas, NV p. 534-539.
18. Wu Y, Zhang R, Hu X, Kashioka H. Learning unsupervised SVM classifier for answer selection in web question answering. 2007 Jun 28 Presented at: Proceedings of the joint conference on empirical methods in natural language processing and computational natural language learning; June 28-30, 2007; Prague, Czech Republic p. 33-41.
19. Toba H, Ming Z, Adriani M, Chua T. Discovering high quality answers in community question answering archives using a hierarchy of classifiers. Inform Sciences 2014:101-115.
20. Shah C, Pomerantz J. Evaluating and predicting answer quality in community QA. 2010 Jul 19 Presented at: SIGIR : Proceedings of the 33rd annual international ACM SIGIR conference on research development in information retrieval; July 19-23, 2010; Switzerland p. 411-418.
21. Arai K, Handayani A. Predicting quality of answer in collaborative Q/A community. International journal of advanced research in artificial intelligence 2013;2(3):21-25.
22. Shah C, Kitzie V, Choi E. Questioning the question - addressing the answerability of questions in community question-answering. 2014 Jan 06 Presented at: 47th Hawaii international conference on system sciences; January 6-9, 2014; Waikoloa, HI p. 1386-1395.
23. Tian Q, Zhang P, Li B. Towards predicting the best answers in community-based question-answering services. 2013 Jul 08 Presented at: Proceedings of the 7th international conference on weblogs and social media; July 8-11, 2013; Cambridge, MA p. 725-728.
24. Athenikos SJ, Han H. Biomedical question answering: A survey. Comput Methods Programs Biomed 2010 Jul;99(1):1-24. [doi: 10.1016/j.cmpb.2009.10.003] [Medline: 19913938]
25. Morris M, Teevan J, Panovich K. What do people ask their social networks, and why?: a survey study of status message q&a behavior. 2010 Apr 10 Presented at: Proceedings of the SIGCHI conference on human factors in computing systems; April 10-15, 2010; Atlanta, GA p. 1739-1748.
26. Luo J, Zhang G, Wentz S, Cui L, Xu R. SimQ: Real-time retrieval of similar consumer health questions. J Med Internet Res 2015;17(2):e43 [FREE Full text] [doi: 10.2196/jmir.3388] [Medline: 25689608]
27. Morris TA, Guard JR, Marine SA, Schick L, Haag D, Tsipis G, et al. Approaching equity in consumer health information delivery: NetWellness. J Am Med Inform Assoc 1997;4(1):6-13 [FREE Full text] [Medline: 8988468]
28. Wong W, Thangarajah J, Padgham L. Contextual question answering for the health domain. J Am Soc Inf Sci Tec Nov 2012;63(11):2313-2327.
29. Nigam K, McCallum A, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. Machine learning 2000;39(2-3):103-134.

30. Liu XY, Zhou YM, Zheng RS. Sentence similarity based on dynamic time warping. 2007 Sep 17 Presented at: International conference on semantic computing, proceedings; September 17-19, 2007; Irvine, CA p. 250-256.
31. Zobel J, Moffat A. Exploring the similarity space. ACM SIGIR Forum 1998;32(1):18-34.
32. Levenshtein V. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 1966;10(8):707-710.
33. Wagner R, Fischer M. The string-to-string correction problem. Journal of ACM 1974;21(1):168-173.
34. Aronson A. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. 2001 Presented at: Proceedings / AMIA annual symposium; 2001; Washington, DC p. 17-21.
35. Wongchaisuwat P, Klabjan D, Jonnalagadda S. GitHub. 2015. A semi-supervised learning approach to enhance community-based question answering: Code and dataset URL: https://github.com/papisw/Health-QA [accessed 2016-07-06] [WebCite Cache ID 6inyvFKJ3]

Abbreviations

CQA: community-based question answering
DTW: Dynamic time warping
EM: expectation maximization
MRR: mean reciprocal rank
NNET: neural networks
QA: questions and answers
SVM: support vector machine
UMLS: Unified medical language system

Edited by G Eysenbach; submitted 05.02.16; peer-reviewed by C Giraud-Carrier, H Zhai; comments to author 22.02.16; revised version received 05.05.16; accepted 24.06.16; published 02.08.16.

Please cite as:
Wongchaisuwat P, Klabjan D, Jonnalagadda SR
A Semi-Supervised Learning Approach to Enhance Health Care Community–Based Question Answering: A Case Study in Alcoholism
JMIR Med Inform 2016;4(3):e24
URL: http://medinform.jmir.org/2016/3/e24/
doi:10.2196/medinform.5490
PMID:27485666

Original Paper

Characterizing the (Perceived) Newsworthiness of Health Science Articles: A Data-Driven Approach

Ye Zhang[1], MS; Erin Willis[2], PhD; Michael J Paul[3], PhD; Noémie Elhadad[4], PhD; Byron C Wallace[5], PhD

[1]Department of Computer Science, University of Texas at Austin, Austin, TX, United States

[2]College of Media, Communication and Information, University of Colorado Boulder, Boulder, CO, United States

[3]Department of Information Science, University of Colorado Boulder, Boulder, CO, United States

[4]Biomedical Informatics, Columbia University, New York, NY, United States

[5]College of Computer and Information Science, Northeastern University, Boston, MA, United States

Corresponding Author:
Ye Zhang, MS
Department of Computer Science
University of Texas at Austin
Room 5.520, 1616 Guadalupe St, Austin, TX, 78701
Austin, TX, 78701
United States
Phone: 1 4127360156
Fax: 1 512 471 3971
Email: yezhang1989@gmail.com

Abstract

Background: Health science findings are primarily disseminated through manuscript publications. Information subsidies are used to communicate newsworthy findings to journalists in an effort to earn mass media coverage and further disseminate health science research to mass audiences. Journal editors and news journalists then select which news stories receive coverage and thus public attention.

Objective: This study aims to identify attributes of published health science articles that correlate with (1) journal editor issuance of press releases and (2) mainstream media coverage.

Methods: We constructed four novel datasets to identify factors that correlate with press release issuance and media coverage. These corpora include thousands of published articles, subsets of which received press release or mainstream media coverage. We used statistical machine learning methods to identify correlations between words in the science abstracts and press release issuance and media coverage. Further, we used a topic modeling-based machine learning approach to uncover latent topics predictive of the perceived newsworthiness of science articles.

Results: Both press release issuance for, and media coverage of, health science articles are predictable from corresponding journal article content. For the former task, we achieved average areas under the curve (AUCs) of 0.666 (SD 0.019) and 0.882 (SD 0.018) on two separate datasets, comprising 3024 and 10,760 articles, respectively. For the latter task, models realized mean AUCs of 0.591 (SD 0.044) and 0.783 (SD 0.022) on two datasets—in this case containing 422 and 28,910 pairs, respectively. We reported most-predictive words and topics for press release or news coverage.

Conclusions: We have presented a novel data-driven characterization of content that renders health science "newsworthy." The analysis provides new insights into the news coverage selection process. For example, it appears epidemiological papers concerning common behaviors (eg, alcohol consumption) tend to receive media attention.

(JMIR Med Inform 2016;4(3):e27) doi:10.2196/medinform.5353

KEYWORDS

natural language processing; text classification; press release; media coverage

Introduction

Background

Health news is an increasingly popular topic in news media [1] and has been shown to improve health outcomes [2,3]. Communicating health science in layman's terms can often be difficult. Information subsidies, such as press releases, are resources for journalists that mitigate this difficulty by facilitating information transfer. The role of information subsidies and their importance to the development of health news and agenda building is related to the demands of the journalism industry [4].

Gandy [5] first defined *information subsidy* as source information provided to a newsroom, and Berkowitz and Adams [6] further defined subsidy as anything provided to the media in order to gain time or space. Press releases, which are often written by journal staff members in the form of news stories, are one type of information subsidy. To increase the rate of publication, public relations practitioners write press releases with journalistic news values, defined as the elements of a story that make it likely to be published [7]. News values, such as proximity, significance, and novelty, act as criteria for deciding what is newsworthy and most likely to increase audience attention.

In this study, we aim to use data-driven, quantitative approaches to address the following questions: What topical content in health science articles correlates with receiving, or not receiving, a press release? Relatedly, what topical content correlates with receiving, or not receiving, news media coverage? What are the differences in the content of articles covered by the news media versus those that receive a press release?

Motivation and Related Work

The news media are powerful conduits by which to disseminate important information to the public [8]. There is a chasm between the constant demand for up-to-date information and shrinking budgets and staff at newspapers around the globe. Information subsidies such as press releases are often looked to as a way to fill this widening gap. As a standard of industry practice, public relations professionals generate packaged information to promote their organization and to communicate aspects of interest to target the public [9].

Agenda setting has been used to explain the impact of the news media in the formation of public opinion [10]. The theory posits that the decisions made by news gatekeepers (eg, editors and journalists) in choosing and reporting news plays an important part in shaping the public's reality. Information subsidies are tools for public relations practitioners to use to participate in the building process of the news media agenda [11,12].

In the area of health, journalists rely more heavily on sources and experts because of the technical nature of the information [12,13]. Tanner [14] found that television health-news journalists reported relying most heavily on public relations practitioners for story ideas. Another study of science journalists at large newspapers revealed that they work through public relations practitioners and also rely on scientific journals for news of medical discoveries [15]. Viswanath and colleagues [4] found that health and medical reporters and editors from small media organizations were less likely to use government websites or scientific journals as resources, but were more likely to use press releases. In other studies, factors such as newspaper circulation, publication frequency, and community size were shown to influence publication of health information subsidies [16-18].

This study focuses on media coverage of developments in health science and scientific findings. Previous research has highlighted factors that might promote press release generation for, and news coverage of, health science articles. This work has relied predominantly on qualitative approaches. For instance, Woloshin and Schwartz [19] studied the press release process by interviewing journal editors about the process of selecting articles for which to generate press releases. They also analyzed the fraction of press releases that reported study limitations and related characteristics. Tsfati et al [20] argued through content analysis that scholars' beliefs in the influence of media increases their motivation and efforts to obtain media coverage, in turn influencing the actual amount of media coverage of their research.

In this study, we present a complementary approach using data-driven, quantitative methods to uncover the topical content that correlates with both news release generation and mainstream media coverage. Our hypothesis is that there exist specific topics—for which words and phrases are proxies—that are more likely to be considered "newsworthy." Identifying such topics will illuminate latent biases in the journalistic process of selecting scientific articles for media coverage.

Contributions

In this work, we apply natural language processing and statistical machine learning techniques to characterize features of scientific articles that receive media coverage. Specifically, we aim to build interpretable statistical models that can reliably predict whether a published health science article will (1) receive a press release from the publishing journal and (2) garner media coverage in mainstream outlets.

To explore these processes empirically we have constructed novel datasets. Our preliminary work [21] showed that one can induce models to reliably discriminate between articles that receive press coverage and those that do not using "bag-of-words" representations of articles with count variables for unigrams and bigrams extracted from article titles and abstracts—unigrams are single words and bigrams are sequences of two adjacent words. Here we substantially extend this preliminary work as follows:

1. We use supervised latent Dirichlet allocation (sLDA) [22] to uncover discriminative topics that correlate with media attention, in addition to simple n-gram correlations.

2. We analyze a new corpus [23] that contains information concerning both press release issuance and media coverage for all articles it contains. Press releases were issued for all articles in this set, but only a subset garnered media attention, thus providing opportunity to disentangle factors that correlate with each type of press.

Our models are able to reliably discriminate between articles that will and will not (1) motivate a press release and (2) receive media coverage. We report robust predictors for these two tasks, both in terms of words and bigrams in a discriminative bag-of-words framework and with respect to higher-level topics uncovered via sLDA.

Methods

Datasets

We now describe the datasets that we have constructed to empirically investigate patterns in press release generation for, media coverage of, and social media attention to, health science articles. We made all of these datasets publicly available, along with our code, to facilitate future research [24].

First, we augmented the dataset recently introduced by Sumner and colleagues [23] in their work addressing the association between exaggeration in health-related science news articles and academic press releases. We will refer to this dataset as Sumner. It contains 462 press releases written for articles published in biomedical and health-related journals by 20 leading UK universities in 2011. For each press release, the authors sourced the corresponding journal article and print or online news stories from national press outlets using the Nexis database, the BBC, Reuters, and Google; the number of news stories per press release ranged from 0 to 10.

Sumner and colleagues coded each journal article, press release, and news piece using a detailed protocol that is available online [25]. We derived two corpora from the Sumner dataset: one was used to investigate press release (PR) issuance, which we call Sumner PR, and the other was used to model news coverage (NC), which we call Sumner NC.

Additionally, we constructed two datasets, Journal of the American Medical Association (JAMA) and Reuters, which we have described in our earlier work [21]. For both of these datasets, we had to generate *negative* instances: health science articles that did not receive media coverage, or for which no press releases were written. To this end, we relied on a novel matched sampling approach [26] aimed at identifying articles that did not garner attention but that had similar characteristics (ie, were published in the same year and in the same journal) to those that did. We describe this process in greater detail below.

We decomposed our aims into distinct modeling tasks to be undertaken using the associated datasets. We treated these as predictive tasks for validation purposes, but our interest is primarily in the predictive features, rather than classifier performance, as such. Table 1 summarizes the four tasks and their corresponding corpora.

Table 1. Summary of the four tasks and their associated datasets.

Task	Source	Positive instances in dataset, n (%)	Negative instances in dataset, n (%)	Title length (words), mean (SD)	Abstract length (words), mean (SD)
PR[a]	Sumner PR (N=3024)	422 (13.96)	2602 (86.04)	13 (5)	214 (67)
PR	JAMA[b] (N=10,760)	846 (7.86)	9914 (92.14)	13 (5)	335 (82)
NC[c]	Sumner NC (N=422)	214 (50.7)	208 (49.3)	14 (5)	226 (79)
NC	Reuters (N=28,910)	1343 (4.65)	27,567 (95.35)	14 (6)	267 (86)

[a]PR: press release.

[b]JAMA: Journal of the American Medical Association.

[c]NC: news coverage.

Press Release Datasets

Sumner Press Release

Our first use of the Sumner corpus [23] involved constructing a dataset to use to induce a discriminative model to predict which scientific articles will receive press releases. To achieve this, we needed to link press releases to the corresponding scientific publications that they cover. For this, we relied on the search functionality in PubMed [27], which provides an interface for searching the over 24 million publications indexed by MEDLINE. We used this to identify the journal articles corresponding to each entry in the Sumner corpus. Specifically, we searched PubMed for the original journal article using the title entered in the coding sheet. In this way, we identified citation information—title, abstract, and Medical Subject Headings (MeSH) keywords—for 422 out of the 460 articles covered by press releases in the Sumner corpus. We were unable to find the remaining 38 articles on PubMed.

All 422 of these articles constitute *positive* examples, because all received press releases. We therefore collected *negative* instances via the matched sampling approach, which proceeded as follows. For each citation, we sampled up to 10 articles from the same journal and the same issue for which no press releases were issued. Our aim in so doing was to isolate content predictors that correlate with garnering media attention, independent of publication venue and temporal factors. In total, we retrieved 2602 citations using this approach. Figure 1 depicts a pair of positive and negative snippets.

Figure 1. A pair of positive and negative instance snippets from the Sumner press release (PR) dataset.

A positive instance snippet
Title: Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis.
MeSH terms: Humans, Child, Preschool,Infant...
Abstract: The global burden of disease attributable to seasonal influenza virus in children is unknown. We aimed to estimate the global incidence of and mortality from lower respiratory infections associated with influenza in children younger than 5 years ... We estimated there were 28,000-111,500 deaths in children younger than 5 years attributable to influenza-associated ALRI in 2008, with 99% of these deaths occurring in developing countries...

A negative instance snippet
Title:Naloxone-precipitated morphine withdrawal behavior and brain IL-1beta expression: comparison of different mouse strains.
MeSH terms: Cells, Animals, Mice, Sequence Analysis, DNA,Neuroglia/physiology ...
Abstract: The development of opioid dependence involves classical neuronal opioid receptor activation ... Analysis of brain nuclei (medial prefrontal cortex, cortex, brain stem, hippocampus, and midbrain and diencephalon regions combined) revealed that, of inbred wild-type mice, there are significant main effects of morphine treatment on IL-1beta expression in the brain regions analyzed ...

Journal of the American Medical Association

The JAMA corpus comprised 846 positive instances, defined as articles for which journal editors created a press release—all journals in this corpus belong to the JAMA network [21]. Negative instances were again selected via matched sampling, focusing on articles from the same journal and year, but for which no press release was issued. After removing duplicates, this corpus comprised 9914 *negative* articles. This collection was exhaustive, containing all press releases available on the JAMA Web archive from October 1, 2012, to October 1, 2014.

News Coverage Datasets

Sumner News Coverage

For the first news coverage prediction task, we used the 422 articles contained in the Sumner dataset. In this case, we knew which articles were covered by one or more news outlets, and we could therefore derive positive and negative labels for each article. In all, 214 of these articles received news media coverage. We will refer to this dataset as Sumner NC.

Reuters

The Reuters corpus [21] comprised health news stories that reported on particular biomedical and health research studies published by the Reuters news agency. In each story, Reuters journalists cited and linked to the original scientific article on which the story reported. Thus, the Reuters stories and their corresponding scientific articles provided us with *positive* instances for the media coverage prediction task. We again used our matched sampling method to sample up to 20 articles for each positive instance as described in Wallace et al [21]. Briefly, we sampled citations published in the same journal, year, and volume as *positive* instances. This resulted in 1343 positive instances and 27,567 negative instances.

Machine Learning Algorithms

Overview

In this section, we describe the machine learning methods we used to analyze the corpora. Broadly, these can be decomposed into our discriminative learning approach and the generative supervised topic modeling method we used to uncover latent topics that correlate with newsworthiness.

Discriminative Learning

For discriminative learning, we used standard logistic regression with a squared $\ell 2$ norm penalty placed on the weights for regularization. Specifically, given a labeled corpus, we optimize the objective in Equation 1 in Figure 2. In Figure 2, X_i is the feature vector representing the i th article—comprising counts of uni- and bigrams— y_i is the label for this article, w is the weight vector to be estimated from the data, and w_0 is an intercept term. We fit this model using LIBLINEAR (Machine Learning Group at National Taiwan University) [28]. λ is a scalar hyper-parameter that controls the trade-off between regularization strength and empirical predictive performance on the training set. We performed five-fold cross-validation and reported average area under the curve (AUC) scores. Cross-validation is a standard means of assessing model performance in which one splits the data into k disjoint "folds" (here k=5) and holds one out at a time. The model is then trained using k-1 folds, and performance metrics are calculated on the held-out fold. This process is repeated k times, resulting in k estimates of performance. Here we used the AUC metric, which is a widely used measure of classifier discriminative performance that captures the probability that a given positive instance will be ranked above an arbitrary negative instance by the model. To select the λ hyper-parameter (Equation 1), we performed a logarithmic line search over possible values ranging from 0.00001 to 100—smaller λ values correspond to stronger regularization. We kept the value that maximized average performance, as assessed via nested cross-validation; thus, we performed λ selection independently for each fold, as this was tuned on the available training data.

As features in the logistic regression model, we used uni- and bigrams extracted from titles, abstracts, and MeSH terms. MeSH terms are Medical Subject Headings drawn from a controlled vocabulary maintained by the National Library of Medicine (NLM). These are manually assigned to citations by trained annotators at the NLM.

For text preprocessing, we used a standard English stop word list, and only kept features that appeared in at least two instances in a given dataset. We kept, at most, the 50,000 most frequently occurring features in the datasets, in cases where there were more than 50,000 unique features. The numbers of features for

each task are summarized together with the sLDA model in the next section.

To identify robustly predictive features, we used bootstrap sampling to construct confidence intervals around coefficient point estimates. Specifically, we fit a regularized logistic regression model to each bootstrap training sample and recorded estimated coefficient values for each feature. We repeated this process 1000 times, deriving a variance from the observed estimates. We then constructed an approximate 95% confidence interval around coefficients using the normal approximation method [29].

Figure 2. Equation 1. X_i is the feature vector representing the *i* th article—comprising counts of uni- and bigrams— y_i is the label for this article, *w* is the weight vector to be estimated from the data, and *w*is an intercept term. λ is a scalar hyper-parameter that controls the trade-off between regularization strength and empirical predictive performance on the training set.

$$\frac{1}{2}w^T w + \lambda \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + w_0)) + 1)$$

Supervised Topic Modeling

Statistical topic models have emerged as an important tool for discovering topics from large collections of text documents. Topic models postulate a *generative story*, in which each document comprises a mixture of topics and each topic corresponds to a probability distribution over words. This is the model specified by latent Dirichlet allocation (LDA) [30].

Supervised topic modeling is a variant of this, in which auxiliary meta-data about documents (ie, supervision) is assumed to be available [31]. Typically, this supervision is expressed as labels or tags on documents. In sLDA, one then assumes a model similar to that of standard LDA: a document is again associated with a distribution over topics that are in turn modeled as distributions over words. However, sLDA extends this to additionally model the document attributes (ie, labels), conditioned on estimated topic frequencies. In our case, the label for a given document was whether or not it received a press release or media coverage—we model these separately. Thus, we aimed to uncover topics that explicitly correlated with press release issuance and media coverage.

More specifically, we assumed that there are *K* topics in the corpus, and the number of class labels is *C*. The model parameters are as follows: the *K* topics $\beta_{1:K}$ (each β_K is a vector of term probabilities), the Dirichlet hyper-parameter α, and a set of prediction coefficients for each class *c*. Each coefficient η_c is a *K*-dimensional vector of real values. The process for generating an article and its label is then modeled as follows:

1. Draw topic proportions θ~Dirichlet(α).

2. For each word in position *n* in the article,

(a) Draw topic assignment $z_n|\theta \sim$ Multinomial (θ)

(b) Draw word as in Figure 3.

3. Draw class label as in Figure 4, where *N* is the total number of words in the article, and the empirical topic frequencies of the article is as shown in Figure 5. The softmax function is as shown in Equation 2 in Figure 6.

Here, the labels *c* for each article are binary: they either received a press release or media coverage, or did not. For parameter estimation, we used the approximate inference algorithm presented in Wang et al [31]. We set the number of topics *K* to 20, which we viewed as an intuitively reasonable number of topics to assume. We set the symmetric Dirichlet prior α to 1.

The words comprising our vocabulary were unique unigrams extracted from citation titles, abstracts, and MeSH terms. We again kept up to 50,000 of the most frequently occurring words in the dataset as features. Ultimately, for the discriminative task—for which we used logistic regression—we used the following: 50,000 features for Sumner PR; 50,000 for JAMA; 10,004 for Sumner NC, which is much smaller; and 50,000 features for the Reuters corpus. For generative modeling (ie, using sLDA), we are left with the following: 23,561 features for the Sumner PR dataset; 23,539 for the JAMA dataset; 5796 for Sumner NC; and 50,000 for the Reuters corpus.

Figure 3. Word distribution. w_n is the word at position n, z_n is the topic at position n, and β_K is a vector of term probabilities.

$$w_n | z_n \sim \text{Multinomial}(\beta_{z_n})$$

Figure 4. Class label. c is class label, η is a K-dimensional vector of real values.

$$c | z_{1:N} \sim \text{softmax}(\overline{z}, \eta)$$

Figure 5. Empirical topic frequencies.

$$\overline{z} = \frac{1}{N} \sum z_n$$

Figure 6. Equation 2: softmax function.

$$p(c|\overline{z},\eta) = \exp(\eta_c^T \overline{z}) / \sum_{l=1}^{C} (\eta_l^T \overline{z})$$

Results

Press Release Issuance

Sumner Press Release

With respect to discriminating between articles that did and did not receive a press release in the Sumner PR dataset, we achieved a mean AUC of 0.666 (SD 0.019; range 0.636-0.720 across five folds of cross-validation), indicating relatively strong predictive performance. We report the top 25 most robustly predictive n-gram features— negative and positive—in Textboxes 1 and Textboxes 2. To extract the features that are consistently correlated with positive instances, we ranked the predictors in descending order according to the lower bound of their corresponding confidence intervals, which were derived via bootstrap estimation as discussed above. Similarly, for negative features, we sorted predictors in ascending order of estimated confidence interval upper bounds.

Figures 7 and 8 show coefficient value distributions, as constructed via the bootstrap, for selected features that positively and negatively correlate with press release issuance for articles in Sumner PR dataset, respectively.

We also present output from a 20-topic sLDA model fit to the Sumner PR dataset in Figure 9. The horizontal axis corresponds to the coefficient of the topic, capturing correlation with press release issuance: topics with larger values, toward the right end of the plot, are therefore correlated with press releases being issued (ie, these topics are more likely to appear in articles that receive press releases). We report the top 10 most probable words estimated for each topic. Here we used whether or not an article received a press release as a label.

Textbox 1. Top 25 negative features of press release prediction on Sumner press release dataset, ranked by upper bound of confidence interval.

Top 25 negative features:

- decreased
- study
- results
- sexual
- mice
- based
- research
- signaling
- evaluated
- regarding
- proposed
- protein
- states
- discuss
- program
- various
- lesions
- review
- thyroid
- analyzed
- performed
- overexpression
- medical
- asd

Textbox 2. Top 25 positive features of press release prediction on Sumner press release dataset, ranked by lower bound of confidence interval.

Top 25 positive features:

- uk
- england
- infection
- death
- MH-great-britain (MH prefix indicates a Medical Subject Headings [MeSH] term)
- magnetic
- alcohol
- life
- weeks
- MH-england/epidemiology
- magnetic resonance
- setting
- main outcome
- pregnancy
- british
- functional magnetic
- increase
- resonance imaging
- MH-great-britain MH-humans
- resonance
- TI-study (TI prefix indicates a title term)
- adjusted
- brain
- outcomes

Figure 7. Density curve of coefficient values of four positively predictive words on Sumner press release datasets.

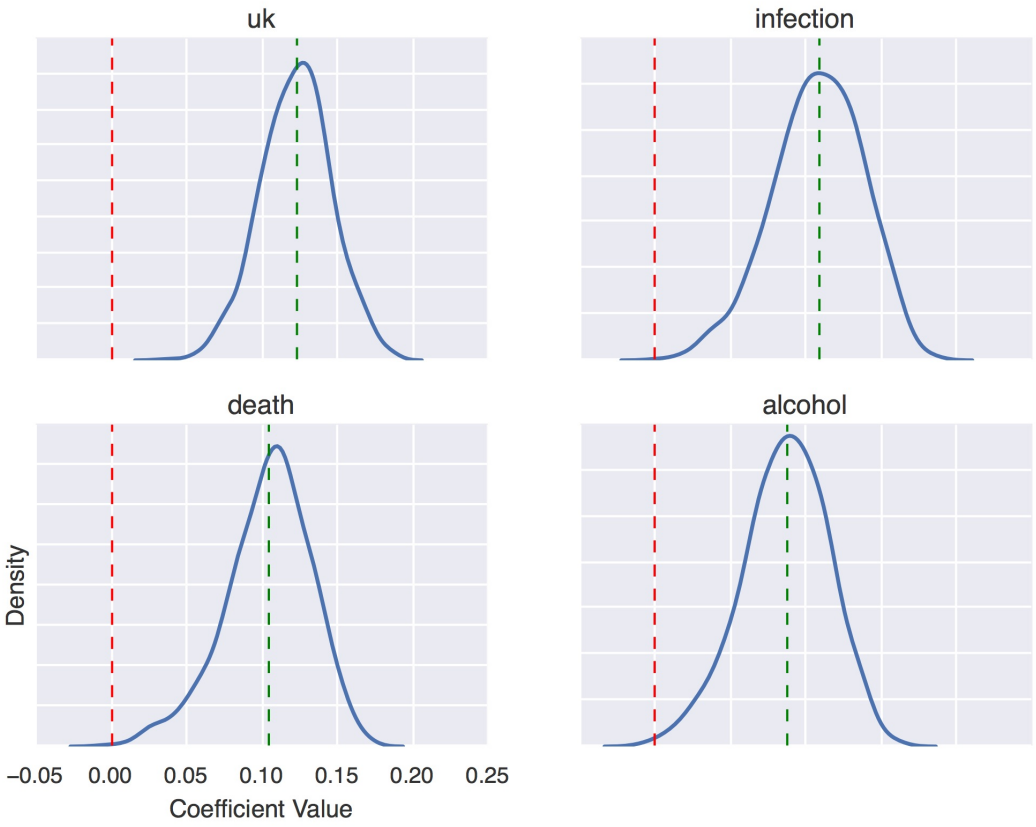


Figure 8. Density curve of coefficient values of four negatively predictive words on Sumner press release datasets.

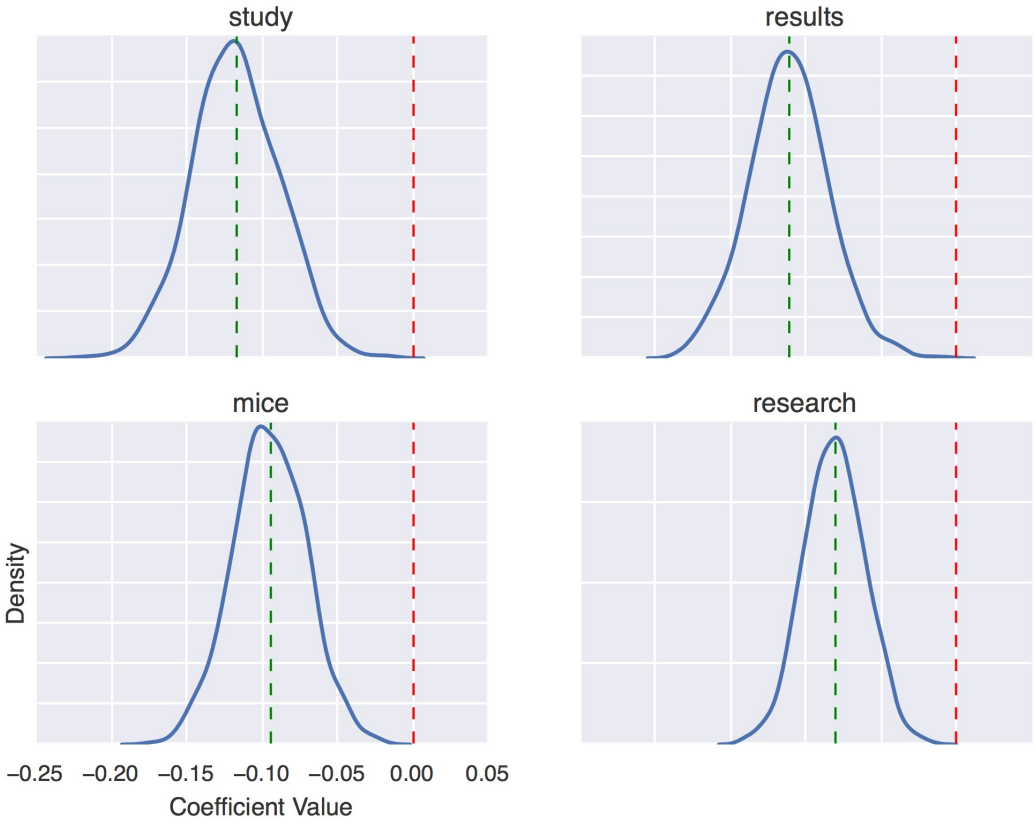


Figure 9. The top 10 most probable words under the topics uncovered by a 20-topic supervised latent Dirichlet allocation model fit to the Sumner press release dataset. The horizontal axis corresponds to the coefficient of the topic. mh: this prefix indicates a Medical Subject Headings (MeSH) term; ti: this prefix indicates a title term.

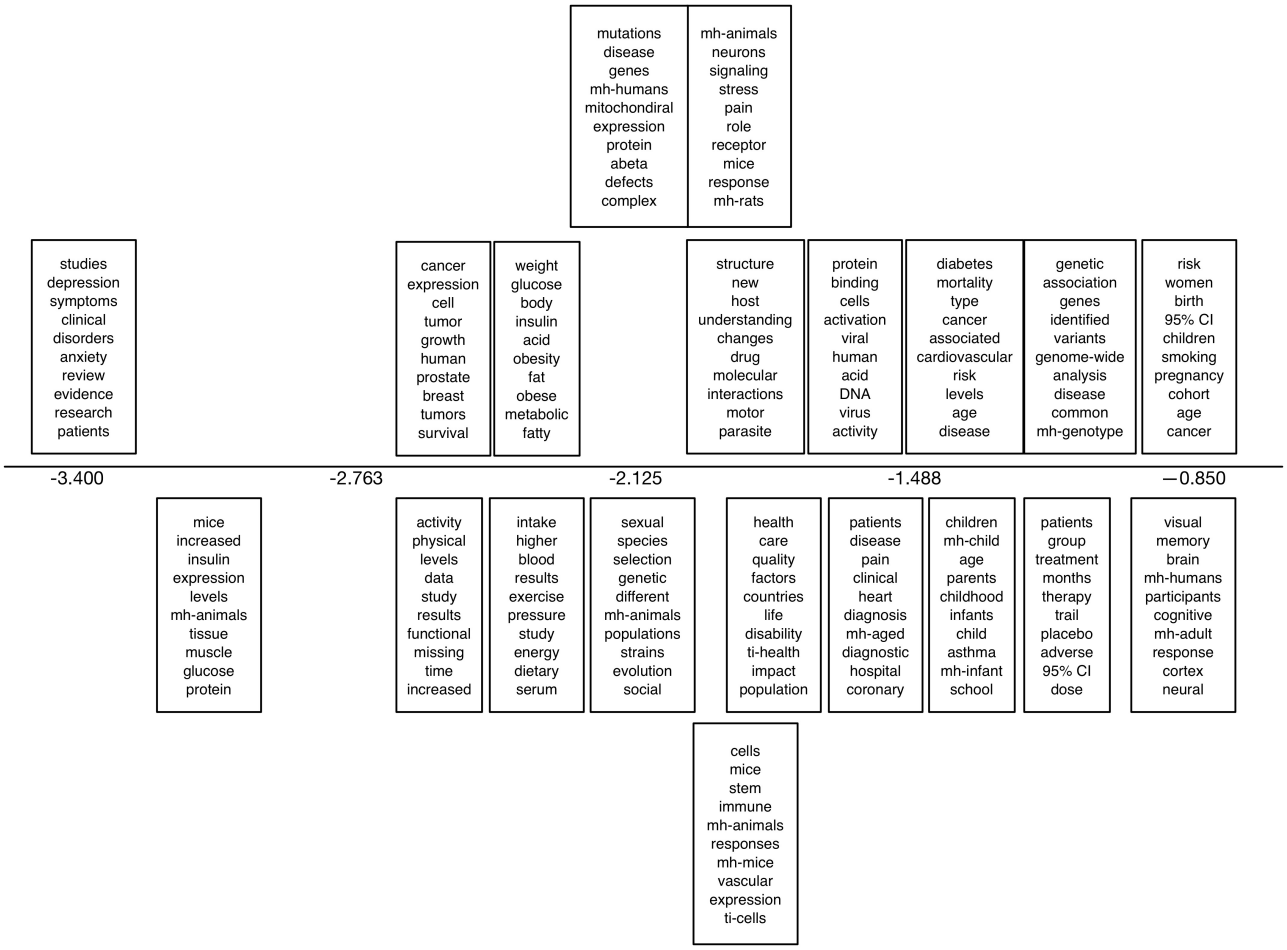
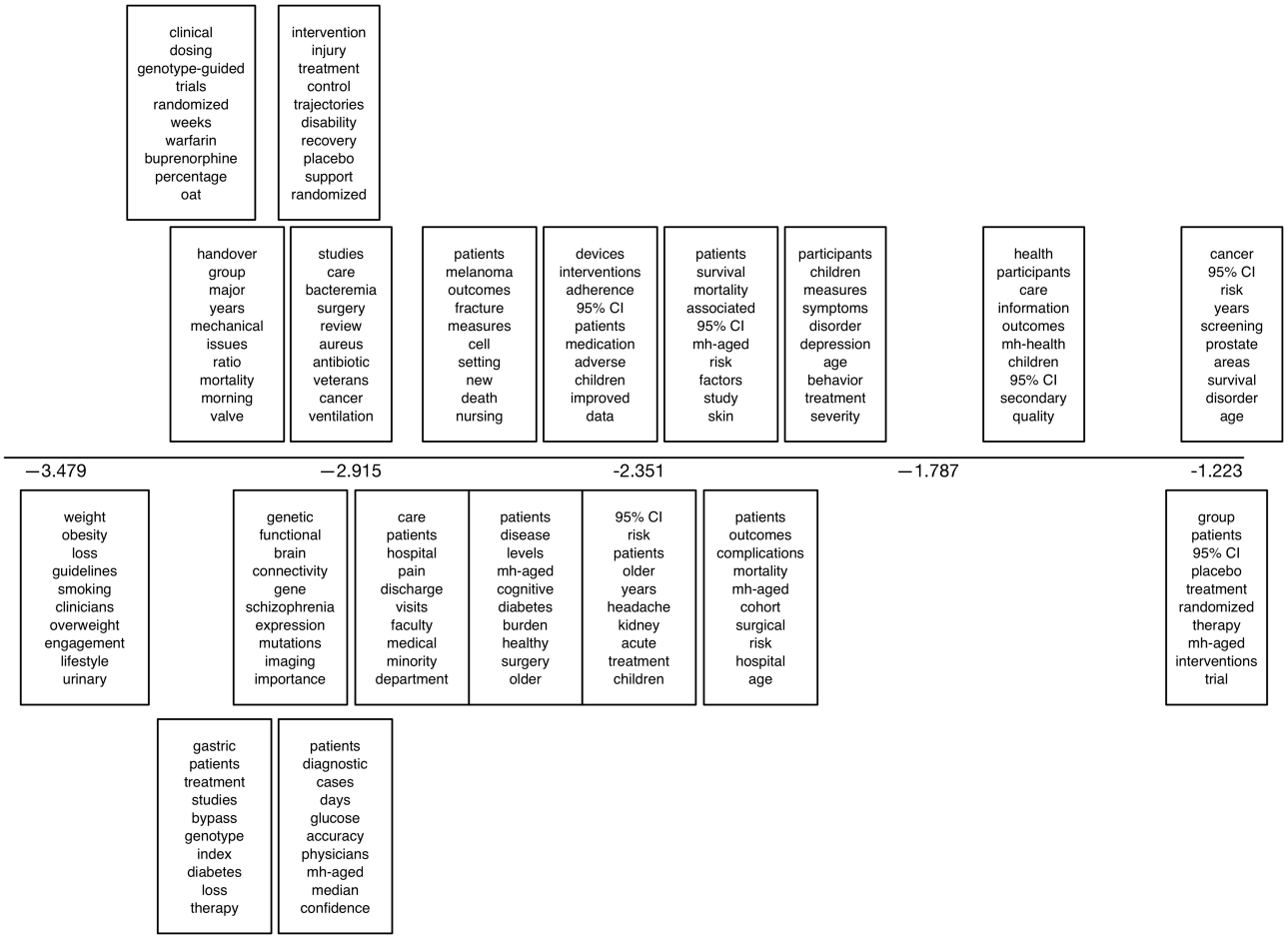


Figure 10. The top 10 most probable words under the topics uncovered by a 20-topic supervised latent Dirichlet allocation model fit to the Journal of the American Medical Association dataset, using press release issuance as the supervision. The horizontal axis corresponds to the coefficient of the topic. mh: this prefix indicates a Medical Subject Headings (MeSH) term.



Journal of the American Medical Association

The analysis reporting informative features for logistic regression prediction was presented in our preliminary work [21], so we do not repeat this here. However, we note that the mean AUC score attained on this dataset was 0.882 (SD 0.018; range 0.853-0.918 across five folds). Figure 10 shows the 20-topic sLDA model fit to this dataset, again using press release issuance as the supervision.

News Coverage Results

Sumner News Coverage

On the Sumner NC dataset, we experimented with two different feature sets: (1) features extracted from the journal articles and (2) features extracted from the corresponding press release text. Our model using journal features achieved a mean AUC of 0.591 (SD 0.044) and ranged from 0.502 to 0.701 across five folds; our model using press release features achieved a mean AUC of 0.575 (SD 0.023) and ranged from 0.497 to 0.622. We note that this exhibits weaker correlation than press release prediction, although it is still better than chance (ie, 0.5). We report the top 25 most predictive features (ie, terms) of news coverage for each feature set in Textboxes 3-6. We rank the features using the same method as in Textboxes 1 and Textboxes 2. In Figures 11 and 12, we show the density curves of coefficient values of four positively predictive words and four negatively predictive words, respectively.

Textbox 3. Top 25 negative features for news coverage prediction on the Sumner news coverage (NC) corpus using the original article abstracts/titles as features.

Top 25 negative article features:

- binding
- receptor
- development
- protein
- resistance
- identify
- surface
- MH-molecular-sequence-data (MH prefix indicates a Medical Subject Headings [MeSH] term)
- direct
- responses
- disruption
- rapidly
- domain
- regions
- structure
- synaptic
- TI-early (TI prefix indicates a title term)
- MH-models-biological
- specific
- using
- culture
- MH-amino-acid-sequence
- TI-children
- understanding
- complexes

Textbox 4. Top 25 negative features for news coverage prediction on the Sumner news coverage (NC) corpus using the original press releases as features.

Top 25 negative press release features:

- resistance
- physical
- proteins
- childhood
- liverpool
- understanding
- born
- impact
- design
- opportunities
- university
- attention
- american
- bristol published
- sentences
- changes
- university birmingham
- discovered
- options
- published
- cardiac
- revealed
- date
- leave
- led professor

Textbox 5. Top 25 positive features for news coverage prediction on the Sumner news coverage (NC) corpus using the original article abstracts/titles as features.

Top 25 positive article features:

- 95% ci
- women
- use
- risk
- countries
- variation
- hazard
- body
- england
- participants
- TI-cohort (TI prefix indicates a title term)
- council
- TI-cohort TI-study
- research council
- medical research
- individual
- individual
- sex
- main outcome
- cohort study
- systolic blood
- relevant
- cancers
- research
- TI-gene

Textbox 6. Top 25 positive features for news coverage prediction on the Sumner news coverage (NC) corpus using the original press releases as features.

Top 25 positive press release features:

- face
- research funded
- shown
- better
- motor
- years
- gene
- england
- edinburgh
- trial
- production
- flu
- council
- targeting
- producing
- roslin
- research council
- roslin institute
- widespread
- royal society
- faces
- providing
- website
- affecting
- exciting

Figure 11. Density curve of coefficient values of four positively predictive words on Sumner news coverage (NC) dataset.

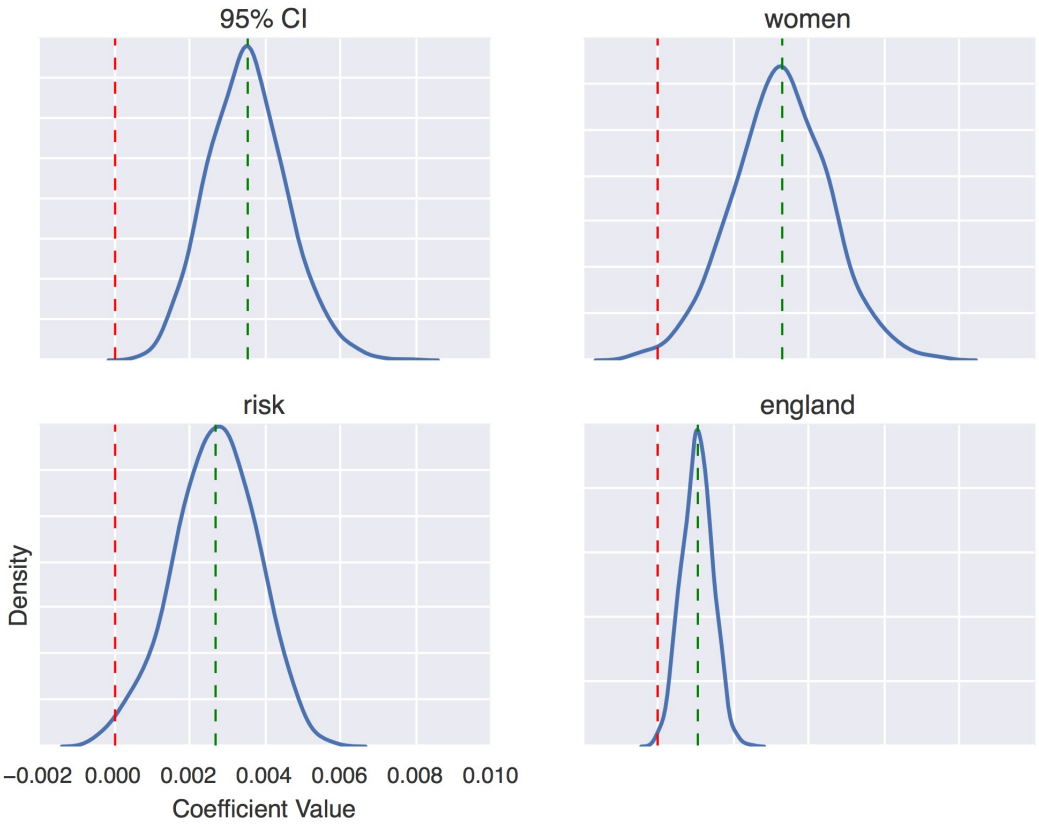


Figure 12. Density curve of coefficient values of four negatively predictive words on Sumner news coverage (NC) dataset.

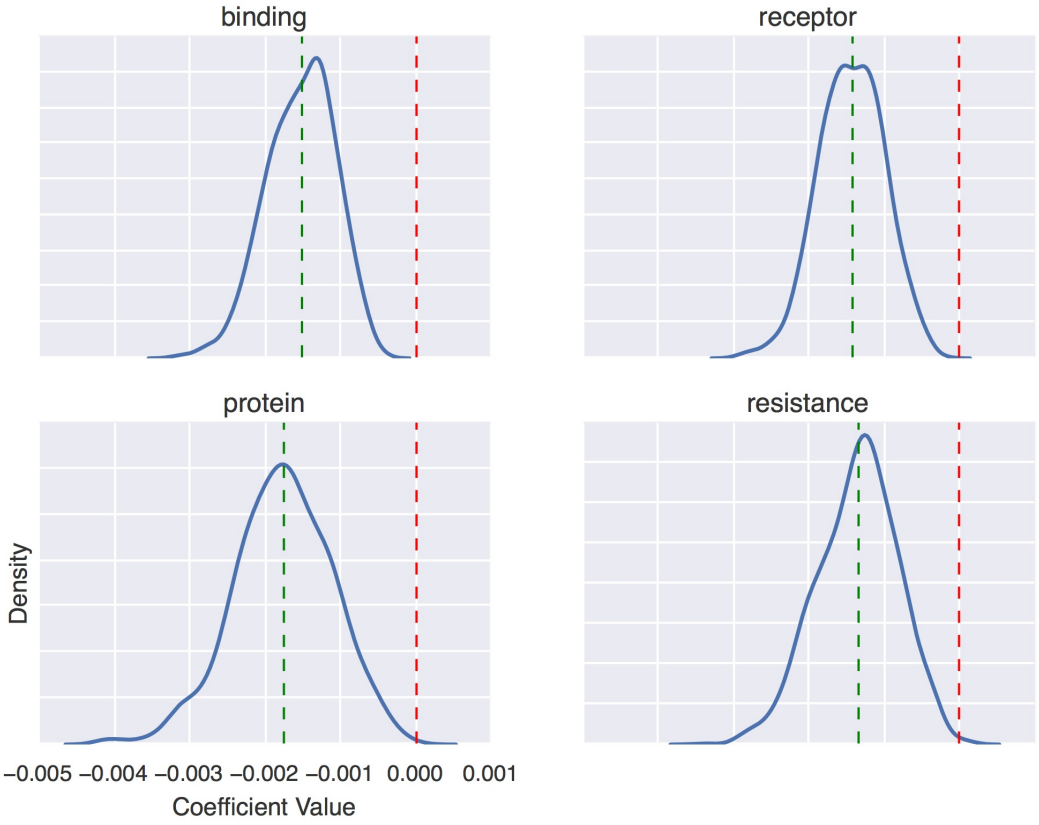
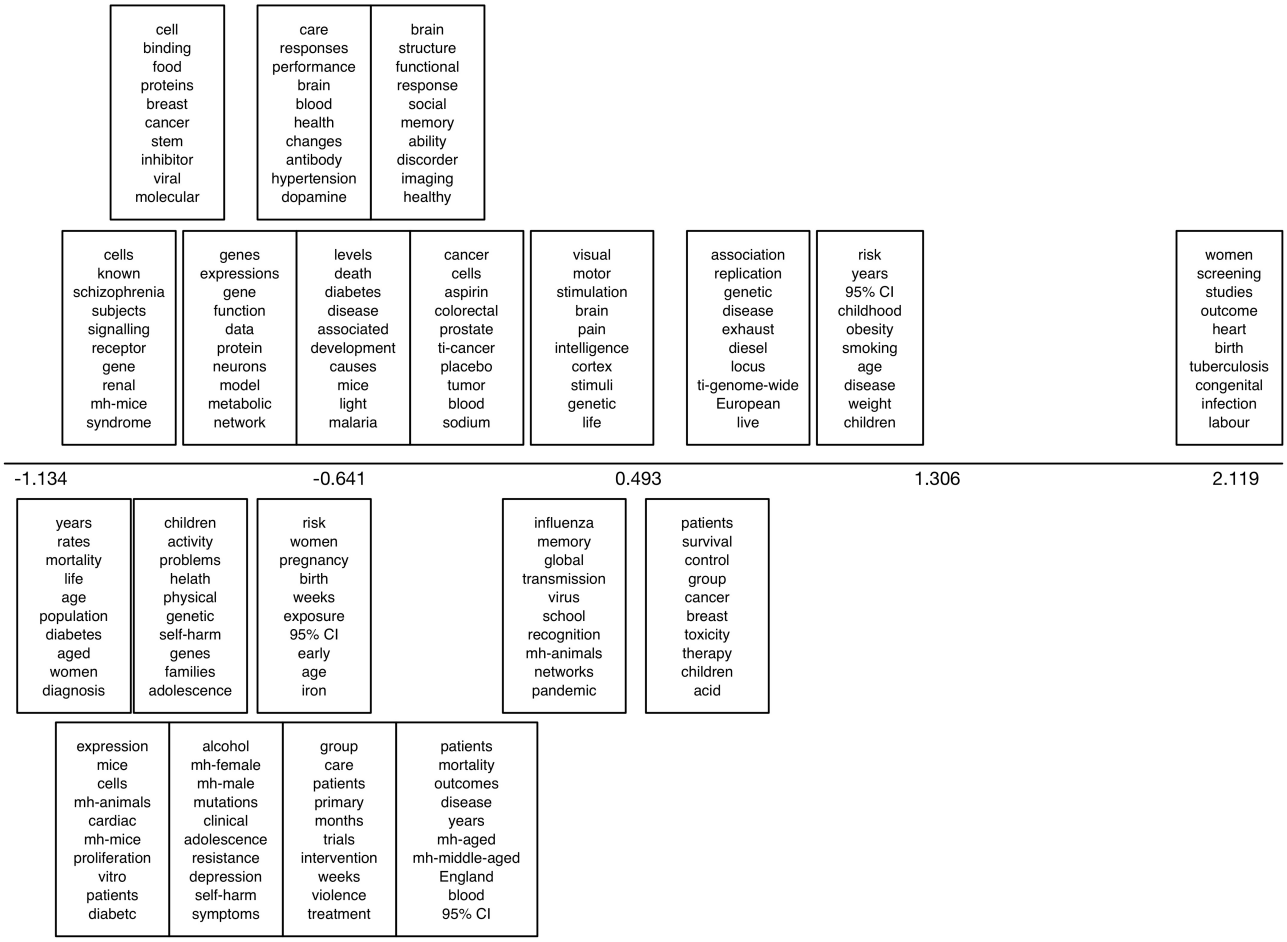


Figure 13. Top 10 most probable words in the topics uncovered by the supervised latent Dirichlet allocation model—again assuming 20 topics—fit to the Sumner news coverage dataset. mh: this prefix indicates a Medical Subject Headings (MeSH) term; ti: this prefix indicates a title term.

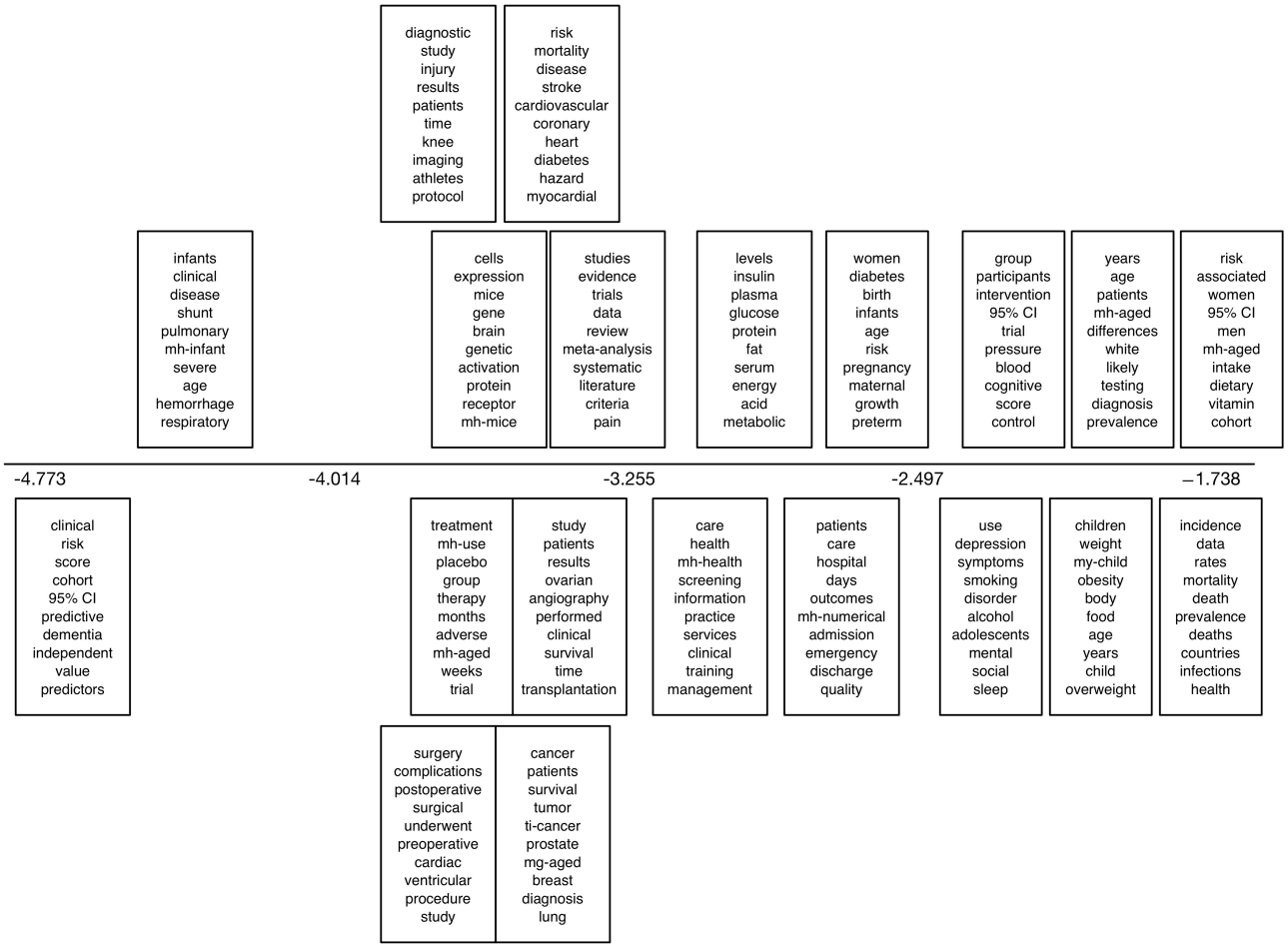


In Figure 13, we show results from the sLDA model fit to the Sumner NC corpus using journal features. The plot is as described above, only the horizontal axis now captures the relative correlation with news coverage estimated for each topic. Here, the supervision captured whether articles received news coverage or not and, hence, we can see which topics (anti)correlate with this.

Reuters

For the discriminative learning task for this dataset, we have reported results previously [21] and do not repeat them here. Briefly, the mean AUC achieved for this task was 0.783 (SD 0.022; range 0.746-0.811). In Figure 14, we report output from the sLDA model: uncovered topics and their degree of correlation with news media coverage. Inspecting the topics suggests a fair amount of overlap with the discriminative topics in Figure 7.

Figure 14. Top 10 words from the 20 topics uncovered by the supervised latent Dirichlet allocation on the Reuters corpus, again using news coverage as the supervision. mh: this prefix indicates a Medical Subject Headings (MeSH) term; ti: this prefix indicates a title term.



Discussion

As news organizations weather the fast-changing information landscape, press releases are now assisting journalists to fill news holes, more than ever before [32]. News organizations are increasingly eschewing the use of specialist reporters such as health science journalists, opting instead to rely on sources and experts [8]. This shift has enabled companies and organizations to play a role in setting the news agenda.

In our prior preliminary work [21], we reported that words such as *women*, *95% CI*, and *drinking* were predictive of both press release issuance and media coverage. Presumably, this reflects interest in population-level results that relate to issues of common concern. Meanwhile, features anticorrelated with press release generation and media coverage seem to be indicative of basic sciences work (eg, binding, receptor, and mice).

This study examined the topics covered by press releases generated by scientific journals. Specifically, we have presented new corpora, methods, and results that aim to illuminate factors that correlate with press release generation for, and news media coverage of, health science articles. Our analysis indicates that scientific journals intentionally disseminate press releases that cover topics likely to be found “newsworthy” by lay audiences. For example, the flu was a topic frequently found in articles deemed newsworthy and in those for which journal editors wrote press releases.

Some of the press release topics were very general and applicable to broad audiences. For example, *women* was a word found frequently in articles that received press releases; indeed, *pregnancy* and *women* were among the most probable words under the topic most strongly correlated with press release issuance (see Figure 4). It is intuitive that most audiences would be interested in research related to women, and pregnancy specifically. By selecting topics that are relevant and applicable to general audiences, scientific journals are helping journalists build the news agenda and educate audiences on (sometimes) difficult and complex topics. Scientific journals are selecting specific studies assumed to be newsworthy by the gatekeepers of the news media, working to form public opinion about a topic. Furthermore, because scientific research is often quite complex, scientific journals may be selecting research studies that are both relatable and easier to translate to a lay audience.

There are several practical implications regarding the results of this study. For instance, press releases from scientific journals might be considered a trustworthy source for journalists working in health news. However, journalists should be aware of the limited scope of the breadth of topics covered in press releases and that other research findings should be explored for news coverage.

This research is not without limitations, however. We can only surmise as to why press releases were written on certain health science research findings or why a press release garnered news

coverage. More research needs to be conducted on why certain health science articles are chosen as newsworthy and why journalists reported on the research findings they did. Although news values are meant to guide journalists' selection of news, there are some who argue that news values are broad and vary greatly among news organizations [33]. Based on the methodology used in this paper, it is also a limitation that no further insight was gleaned from the specific press releases or that the media coverage was not examined. More research must be conducted on how health science findings are being explained in press releases, and how media are translating the press releases into news stories.

Moving forward, we are encouraged by our positive results, and believe our models could be improved further in future work. For example, we could move beyond simple lexical features like n-grams and MeSH terms, including high-level concepts as features, such as the size and composition of the study cohort or the affected population, the type of study (eg, observational or controlled), and whether the research is basic or more applied. Richer linguistic features would also be interesting to incorporate, to help understand if certain writing styles are associated with more or less press coverage. When predicting media coverage, it would also be interesting to use features extracted from the press releases in addition to, or instead of, the features from the original journal articles, to understand how press releases influence the news media.

Conflicts of Interest

None declared.

References

1. Pew Research Center: Journalism and Media Staff. Pew Research Center. Washington, DC: Pew Internet & American Life Project; 2008 Nov 24. Health news coverage in the US media URL: http://www.journalism.org/2008/11/24/health-news-coverage-in-the-u-s-media/ [accessed 2016-09-13] [WebCite Cache ID 6iZB559tI]
2. Dutta-Bergman MJ. Health attitudes, health cognitions, and health behaviors among Internet health information seekers: Population-based survey. J Med Internet Res 2004 May 28;6(2):e15 [FREE Full text] [doi: 10.2196/jmir.6.2.e15] [Medline: 15249264]
3. Shim M, Kelly B, Hornik R. Cancer information scanning and seeking behavior is associated with knowledge, lifestyle choices, and screening. J Health Commun 2006;11 Suppl 1:157-172. [doi: 10.1080/10810730600637475] [Medline: 16641081]
4. Viswanath K, Blake KD, Meissner HI, Saiontz NG, Mull C, Freeman CS, et al. Occupational practices and the making of health news: A national survey of US health and medical science journalists. J Health Commun 2008 Dec;13(8):759-777. [doi: 10.1080/10810730802487430] [Medline: 19051112]
5. Gandy OH. Beyond Agenda Setting: Information Subsidies and Public Policy. Norwood, NJ: Ablex Publishing Corporation; 1982.
6. Berkowitz D, Adams DB. Information subsidy and agenda building in local television news. Journal Mass Commun Q 1990 Dec;67(4):723-731. [doi: 10.1177/107769909006700426]
7. Sallot LM, Steinfatt TM, Salwen MB. Journalists' and public relations practitioners' news values: Perceptions and cross-perceptions. Journal Mass Commun Q 1998 Jun;75(2):366-377. [doi: 10.1177/107769909807500211]
8. McCombs M. Setting the Agenda: The Mass Media and Public Opinion. 2nd edition. Cambridge, UK: Polity Press; 2014.
9. Minnis JH, Pratt CB. Let's revisit the newsroom: What does a weekly newspaper print? Public Relat Q 1995;40(3):13 [FREE Full text]
10. McCombs ME, Shaw DL. The agenda-setting function of mass media. Public Relat Q 1972;36(2):176 [FREE Full text]
11. Curtin PA. Reevaluating public relations information subsidies: Market-driven journalism and agenda-building theory and practice. J Public Relations Res 1999;11(1):53 [FREE Full text] [doi: 10.1207/s1532754xjprr1101_03]
12. Kim JY, Kiousis S. The role of affect in agenda building for public relations implications for public relations outcomes. Journal Mass Commun Q 2012;89(4):657. [doi: 10.1177/1077699012455387]
13. Kruvand M. "Dr. Soundbite": The making of an expert source in science and medical stories. Sci Commun 2012;34(5):566. [doi: 10.1177/1075547011434991]
14. Tanner AH. Agenda building, source selection, and health news at local television stations: A nationwide survey of local television health reporters. Sci Commun 2004;25(4):350. [doi: 10.1177/1075547004265127]
15. Conrad P. Uses of expertise: Sources, quotes, and voice in the reporting of genetics in the news. Public Underst Sci 1999;8(4):285. [doi: 10.1088/0963-6625/8/4/302]
16. Caburnay CA, Luke DA, Cameron GT, Cohen EL, Fu Q, Lai CL, et al. Evaluating the Ozioma cancer news service: A community randomized trial in 24 US cities. Prev Med 2012 Jun;54(6):425-430 [FREE Full text] [doi: 10.1016/j.ypmed.2012.04.010] [Medline: 22546317]
17. Martinson BE, Hindman DB. Building a health promotion agenda in local newspapers. Health Educ Res 2005 Feb;20(1):51-60 [FREE Full text] [doi: 10.1093/her/cyg104] [Medline: 15253997]

18. Morton LP. How newspapers choose the releases they use. Public Relat Rev 1986;12(3):22. [doi: 10.1016/S0363-8111(86)80048-0]
19. Woloshin S, Schwartz LM. Press releases: Translating research into news. JAMA 2002 Jun 5;287(21):2856-2858. [doi: 10.1001/jama.287.21.2856] [Medline: 12038933]
20. Tsfati Y, Cohen J, Gunther AC. The influence of presumed media influence on news about science and scientists. Sci Commun 2010. [doi: 10.1177/1075547010380385]
21. Wallace BC, Paul MJ, Elhadad N. Assoc Adv Artif Intell. 2015. What predicts media coverage of health science articles? URL: http://www.cs.jhu.edu/~mpaul/files/w3phi2015_media.pdf [accessed 2016-09-13] [WebCite Cache ID 6kV9ACbVR]
22. Blei DM, McAuliffe JD. Supervised topic models. In: Proceedings of the 20th International Conference on Neural Information Processing Systems. 2007 Presented at: 20th International Conference on Neural Information Processing Systems; December 3-5, 2007; Vancouver, BC p. 121-128 URL: http://papers.nips.cc/paper/3328-supervised-topic-models.pdf
23. Sumner P, Vivian-Griffiths S, Boivin J, Williams A, Venetis CA, Davies A, et al. The association between exaggeration in health related science news and academic press releases: Retrospective observational study. BMJ 2014 Dec 09;349:g7015 [FREE Full text] [doi: 10.1136/bmj.g7015] [Medline: 25498121]
24. Zhang Y. Github. 2015. A data-driven approach to characterizing the perceived newsworthiness of health science articles URL: https://github.com/yezhang1989/A-Data-Driven-Approach-to-Characterizing-the-Perceived-Newsworthiness-of-Health-ScienceArticles [accessed 2015-11-07] [WebCite Cache ID 6cqz1Mng9]
25. Chambers C. figshare. 2014. InSciOut URL: http://figshare.com/articles/InSciOut/903704 [accessed 2015-11-07] [WebCite Cache ID 6cqzUn1zy]
26. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 1985;39(1):33-38 [FREE Full text]
27. PubMed.gov. URL: http://www.ncbi.nlm.nih.gov/pubmed [WebCite Cache ID 6csE4rYzD]
28. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A library for large linear classification. J Mach Learn Res Jun 2008;9:1871-1874 [FREE Full text]
29. Wasserman L. All of Statistics: A Concise Course in Statistical Inference. New York, NY: Springer Science & Business Media; 2013.
30. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003 Mar;3:993-1022 [FREE Full text]
31. Wang C, Blei D, Li FF. Simultaneous image classification and annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2009 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 20-25, 2009; Miami, FL p. 1903 URL: http://vision.stanford.edu/pdf/WangBleiFei-Fei_CVPR2009.pdf
32. Zoch LM, Supa DW. Dictating the news: Understanding newsworthiness from the journalistic perspective. Public Relat J 2014;8(1) [FREE Full text]
33. Gans HJ. Deciding What's News: A Study of CBS Evening News, NBC Nightly News, Newsweek, and Time. Evanston, IL: Northwestern University Press; 1979:A.

Abbreviations

AUC: area under the curve
JAMA: Journal of the American Medical Association
LDA: latent Dirichlet allocation
MeSH: Medical Subject Headings
MH/mh: MeSH terms
NC: news coverage prediction task
NLM: National Library of Medicine
PR: press release prediction task
sLDA: supervised latent Dirichlet allocation
Sumner NC: Sumner's dataset for news coverage prediction task
Sumner PR: Sumner's dataset for press release prediction task
TI/ti: title terms

Edited by G Eysenbach; submitted 18.11.15; peer-reviewed by J Kimmerle, H Hao, S Jonnalagadda, H Zhai; comments to author 03.03.16; revised version received 02.07.16; accepted 20.07.16; published 22.09.16.

Please cite as:
Zhang Y, Willis E, Paul MJ, Elhadad N, Wallace BC
Characterizing the (Perceived) Newsworthiness of Health Science Articles: A Data-Driven Approach
JMIR Med Inform 2016;4(3):e27
URL: http://medinform.jmir.org/2016/3/e27/
doi:10.2196/medinform.5353
PMID:27658571

Original Paper

Evaluation of an Expert System for the Generation of Speech and Language Therapy Plans

Vladimir Robles-Bykbaev[1,2], MSc, Eng; Martín López-Nores[2], PhD; Jorge García-Duque[2], PhD; José J Pazos-Arias[2], PhD; Daysi Arévalo-Lucero[1], Eng

[1]Grupo de Investigación en Inteligencia Artificial y Tecnologías de Asistencia, Universidad Politécnica Salesiana, Cuenca, Ecuador

[2]AtlantTIC Research Center for Information and Communication Technologies, Department of Telematics Engineering, University of Vigo, Vigo, Spain

Corresponding Author:
Vladimir Robles-Bykbaev, MSc, Eng
Grupo de Investigación en Inteligencia Artificial y Tecnologías de Asistencia
Universidad Politécnica Salesiana
Calle Vieja, 12-30
Elia Liut
Cuenca, 010102
Ecuador
Phone: 593 7862213 ext 1278
Fax: 593 7862213
Email: vrobles@ups.edu.ec

Abstract

Background: Speech and language pathologists (SLPs) deal with a wide spectrum of disorders, arising from many different conditions, that affect voice, speech, language, and swallowing capabilities in different ways. Therefore, the outcomes of Speech and Language Therapy (SLT) are highly dependent on the accurate, consistent, and complete design of personalized therapy plans. However, SLPs often have very limited time to work with their patients and to browse the large (and growing) catalogue of activities and specific exercises that can be put into therapy plans. As a consequence, many plans are suboptimal and fail to address the specific needs of each patient.

Objective: We aimed to evaluate an expert system that automatically generates plans for speech and language therapy, containing semiannual activities in the five areas of hearing, oral structure and function, linguistic formulation, expressive language and articulation, and receptive language. The goal was to assess whether the expert system speeds up the SLPs' work and leads to more accurate, consistent, and complete therapy plans for their patients.

Methods: We examined the evaluation results of the SPELTA expert system in supporting the decision making of 4 SLPs treating children in three special education institutions in Ecuador. The expert system was first trained with data from 117 cases, including medical data; diagnosis for voice, speech, language and swallowing capabilities; and therapy plans created manually by the SLPs. It was then used to automatically generate new therapy plans for 13 new patients. The SLPs were finally asked to evaluate the accuracy, consistency, and completeness of those plans. A four-fold cross-validation experiment was also run on the original corpus of 117 cases in order to assess the significance of the results.

Results: The evaluation showed that 87% of the outputs provided by the SPELTA expert system were considered valid therapy plans for the different areas. The SLPs rated the overall accuracy, consistency, and completeness of the proposed activities with 4.65, 4.6, and 4.6 points (to a maximum of 5), respectively. The ratings for the subplans generated for the areas of hearing, oral structure and function, and linguistic formulation were nearly perfect, whereas the subplans for expressive language and articulation and for receptive language failed to deal properly with some of the subject cases. Overall, the SLPs indicated that over 90% of the subplans generated automatically were "better than" or "as good as" what the SLPs would have created manually if given the average time they can devote to the task. The cross-validation experiment yielded very similar results.

Conclusions: The results show that the SPELTA expert system provides valuable input for SLPs to design proper therapy plans for their patients, in a shorter time and considering a larger set of activities than proceeding manually. The algorithms worked well even in the presence of a sparse corpus, and the evidence suggests that the system will become more reliable as it is trained with more subjects.

(JMIR Med Inform 2016;4(3):e23) doi:10.2196/medinform.5660

KEYWORDS

speech-language pathology; rehabilitation of speech and language disorders; decision support systems, clinical; expert systems

Introduction

Developing and maintaining proper communication skills is a mainstay for every individual to express needs, to learn, to be related with the environment and, in general, to have the opportunity to participate as an active member of society. According to the World Health Organization, individuals with communication difficulties are at a significant social disadvantage in both developing and developed countries [1]. This disadvantage often affects a person's emotional and social life and can compromise educational and job opportunities, particularly in sectors where effective communication is critical, such as health care, education, local government, and justice.

Speech and language therapy (SLT) is an area of health care focused on the evaluation and treatment of a broad range of disorders, which can be roughly classified as affecting voice, speech, language, or swallowing capabilities. Disorders like selective mutism, dysarthria, aphasia, and dysphagia have a substantial impact on quality of life and human potential, whether they affect children who stutter as they struggle to speak up in class, lawyers or teachers with adult-onset voice disorders, or post-stroke individuals laboring to communicate verbally. Numerous studies about the incidence and prevalence of communication disorders in developed countries depict similar realities for Europe, Canada, Australia, and the United States [2-5]. Based on figures like the 7.5 million people in the United States who have voice disorders, the 3 million who stutter, and the 6-8 million who have been diagnosed with some form of language impairment, the American Speech-Language-Hearing Association estimates that 40 million Americans are affected by communication disorders, costing the nation US $154-184 billion annually [6]. It is estimated that more than 60 million people in the European Union are affected, with an estimated cost of €220-260 billion. As the population ages and survival odds improve for fragile infants and individuals who have sustained injury or acquired disease, the number of people with communication disorders will likely continue to increase [7].

Notwithstanding the societal and economic impact of communication disorders, SLT remains a largely overlooked area of health care. The latest World Report on Disability highlights that many countries suffer from lack of professionals, services, and structures to provide effective assessment, diagnosis, counseling, intervention, and treatment for people suffering from communication disorders [8]. In such conditions, speech and language pathologists (SLPs) have very limited time to work with their patients. This may mean that the diagnosis may fail to accurately identify the causes of the disorders, that the designed therapy plans may be suboptimal (eg, because the SLPs fail to keep in mind the whole set of activities they could apply), or that the treatment may be insufficient or not properly applied [7]. In this respect, Turnbull et al [9] found that only 19.2% of young people (from birth to 21 years old) who have communication disorders are actually receiving some form of specific care. Mackenzie et al [10] surveyed SLT provision for people with aphasia in the United Kingdom and found many areas reported low staffing levels and were thus unable to provide the recommended care or a comprehensive service. Code and Heron [11] also concluded that people with aphasia receive significantly less therapy than national recommendations suggest.

Over the last decade, many research efforts have separately shown evidence that the application of information and communication technology (ICT) has great potential to improve the quality and efficiency of SLT practice, as well as health outcomes and patients' quality of life. There have been several approaches to automate diagnostic tests by means of audiovisual signal processing [12-15] and to automate the generation of therapy plans for specific disorders [16,17]. In this paper, we evaluate the support provided to SLPs by the SPELTA (SPEech and Language Therapy Assistant) expert system presented by Robles-Bykbaev et al [18], which aims to automatically generate therapy plans for SLT, containing semiannual activities and daily exercises for an unrestricted range of disorders affecting the five areas of hearing, oral structure and function, linguistic formulation, expressive language and articulation, and receptive language. The goal is to assess whether the expert system can speed up the SLPs' work and lead to more accurate, consistent, and complete therapy plans for their patients.

Methods

SPELTA Expert System

The SPELTA expert system is one part of a set of ICT tools developed by Universidad Politécnica Salesiana (Ecuador) and Universidade de Vigo (Spain) to support SLT within an integrative environment for clinicians and students, pathologists, patients, relatives, and other potential users [19]. The environment is based on a formal knowledge model of the SLT domain and leans on OpenEHR solutions to support the storage and exchange of health-related data. As depicted in Figure 1, the SPELTA system is involved with the automatic generation of therapy plans for new subjects, based on two sources of information: (1) domain ontologies that interrelate the activities and the exercises with specific diseases, speech-language disorders, and skills, and (2) the corpus of patient profiles, containing the compendium of data, plans, and evaluations of previous patients.

Specifically, the profile of an SLT patient contains the following data:

- Personal data, including chronological age, gender, name, etc.
- A medical record specifying diagnosis, general medical conditions and related disorders (eg, cerebral palsy, hemiparesis, athetosis), as indicated by doctors.
- A record of cognitive development data, indicating cognitive age, gap in language development, expressive

language age, and receptive language age (as estimated by SLPs).
- An SLT evaluation that looks at 102 parameters from the five SL areas:

1. Hearing—subjective evaluation of the auditory condition: reflex, localization of sound sources, and response to voice.

2. Oral structure & function—tongue, teeth, palate, lips, and maxillary mobility.

3. Linguistic formulation—phonation and breathing condition.

4. Expressive language and articulation—vocal development, social communication, semantics (content)-vocabulary and concepts, structure (form)-morphology and syntax, and integrative thinking skills; pronunciation of phonemes, sentences, polysyllabic words, and vowel phonemes.

5. Receptive language—attention, semantics (context)-vocabulary and concepts, structure (form)-morphology and syntax, and integration skills.

- A therapy plan, containing five subplans with lists of semiannual activities and daily exercises for each one of the SL areas. One example of activity could be "perform blow exercises to increase the blowing force." Two specific exercises related to this activity could be "blow confetti 10 times during 2 seconds" or "inflate one balloon in no more than 6 exhalations."
- Control evaluations with the results of successive therapy sessions.

Internally, the SPELTA system relies on an implementation of the Partition Around Medoids algorithm to generate clusters of patient profiles with two levels of granularity [20]. The generation of a new therapy plan is dealt with as a classification problem, looking for the most similar cases in each one of the five SL areas according to the K-Nearest Neighbors criterion [21]. First-level clusters represent groups of patients who may have similar speech-language skills and limitations, but possibly arising from (or linked to) different medical conditions. To create these groups, we use the distance metrics of Figure 2, where S_i and S_j refer to two different subjects, A is one of the SL areas, f goes over the set of features from the medical records relevant for that area ($features_{MR}(A)$), and ManhDist denotes the mean-Manhattan binary distance [18].

In the second level, the subjects are clustered according to the fine-grained evaluation of the record of cognitive development data and the initial SLT evaluation. For example, within a first-level cluster that includes the cases of children with Down syndrome and phonological disorders, we need to differentiate subjects who commit additions (ie, adding extra sounds in some words, eg, "balue" for "blue") from subjects who commit substitutions (ie, one or more sounds are substituted for others, eg, "bagon" for "wagon"). In this case, we use the distance metrics of Figure 3. The first summation measures the mean-Manhattan binary distance of the initial SLT evaluations of two subjects, considering only the dimensions relevant to the speech-language area in question, $dimensions_{IE}(A)$. The second summation provides a scale factor derived from the absolute differences of cognitive age, gap in language development, expressive language age, and receptive language age (the features of cognitive development data) [18].

Figure 4 depicts an example of the cluster structure generated by SPELTA for each of the SL areas we consider. Each one of the first-level and second-level clusters has one of the subject cases designated as a medoid, rather than a fictitious case computed by averaging. This facilitates the classification of new cases, identifying the closest subjects in each one of the SL areas.

The plans provided by the SPELTA system are presented to SLPs through visual interfaces, so that they can validate it as a whole or modify certain parts, as they deem necessary. To facilitate the task, the interfaces show which cases were found to be closest in each one of the SL areas. If several subjects were found to be equally distant to the new one in some of the areas, then it is possible to browse the superset of activities, the intersections, and the disjunctions. As an example, Table 1 shows the activities of one master plan generated by the SPELTA system, with the third column indicating the most similar subjects in each area and the features that make them similar to the new case. The profile description is as follows: age 15 years, 8 months; medical diagnosis of athetoid cerebral palsy (ICD-10-CM code G80.3); speech and language diagnosis of mixed receptive-expressive language disorder (ICD-10-CM code F80.1); receptive language age of 4 years; expressive language age of 2 years, 8 months; and a language developmental age of 3 years, 4 months.

Table 1. The activities of a sample therapy plan provided by the SPELTA system (Case 52).

Area	Activities	Source subplans
Hearing	Perform exercises to sounds identification.	Case 37: a patient with a similar receptive language age (4 years, 6 months) and a 100% coincidence in the evaluation of hearing (cochleo-palpebral reflex, startle response, turns head to sound source, identifying sound objects, sound source localization without visual stimulus).
	Discriminate sounds of nature, body, and animals.	
	Perform phonemes discrimination exercises.	
Oral structure & function	Perform segmental relaxation massages.	Case 18: a patient with an 84% coincidence in the oral peripheral mechanism (same tongue size, same speed in tongue movements, present tongue protrusion, voluntary and involuntary swallowing are present, is able to chew hard and soft food, sialorrhea is not present).
	Perform slow and fast tongue movements.	
	Perform exercises with lips (retraction and protrusion).	
	Achieve sound productions using the oropharynx structure.	
	Perform active and passive exercises using tongue, lips, and jaw.	
Linguistic formulation	Work in the automatic respiration process (inspirations and expirations), and work with blow exercises to increase the blowing force.	Case 22: a patient with a 70% coincidence in linguistic formulation (same respiratory frequency, same thorax symmetry, diaphragmatic breathing).
	Respiration exercises associated to vowels and simple phonemes (/pa/, /da/, /fo/).	Case 3: a patient with a 70% coincidence in linguistic formulation (diaphragmatic breathing, no nasal obstruction, same exhalation period).
Expressive language & articulation	Construct sentences from a given word.	Case 22: a patient with a similar expressive language age (1 year, 7 months), similar diagnosis for the medical examination (cerebral palsy and mixed receptive-expressive language disorder) and a 100% coincidence in the speech-language evaluation.
	Sort out the words of a sentence.	
	Work in grammatical structure.	
	Develop the spontaneous conversation	
	Perform activities that use twisters and rhymes.	
	Work with the personal articulation exercise book.	
Receptive language	Work with sequences and puzzles of 4 elements.	Case 37: a patient with a similar receptive language age (4 years, 6 months), similar diagnosis for the medical examination (cerebral palsy and mixed receptive-expressive language disorder) and a 90% coincidence in the speech-language evaluation (the only difference relates to the use of place prepositions like “under,” “over,” etc).
	Learn semantic categories	
	Identify objects according to their utility.	
	Identify daily activities.	
	Learn temporal notions (day and night, before and after).	
	Identify similar/distinct objects according to their utility.	

Figure 1. A block diagram of the SPELTA system.

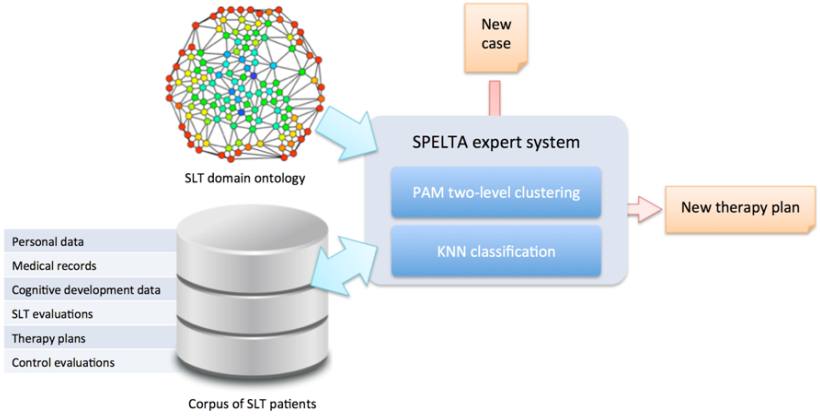


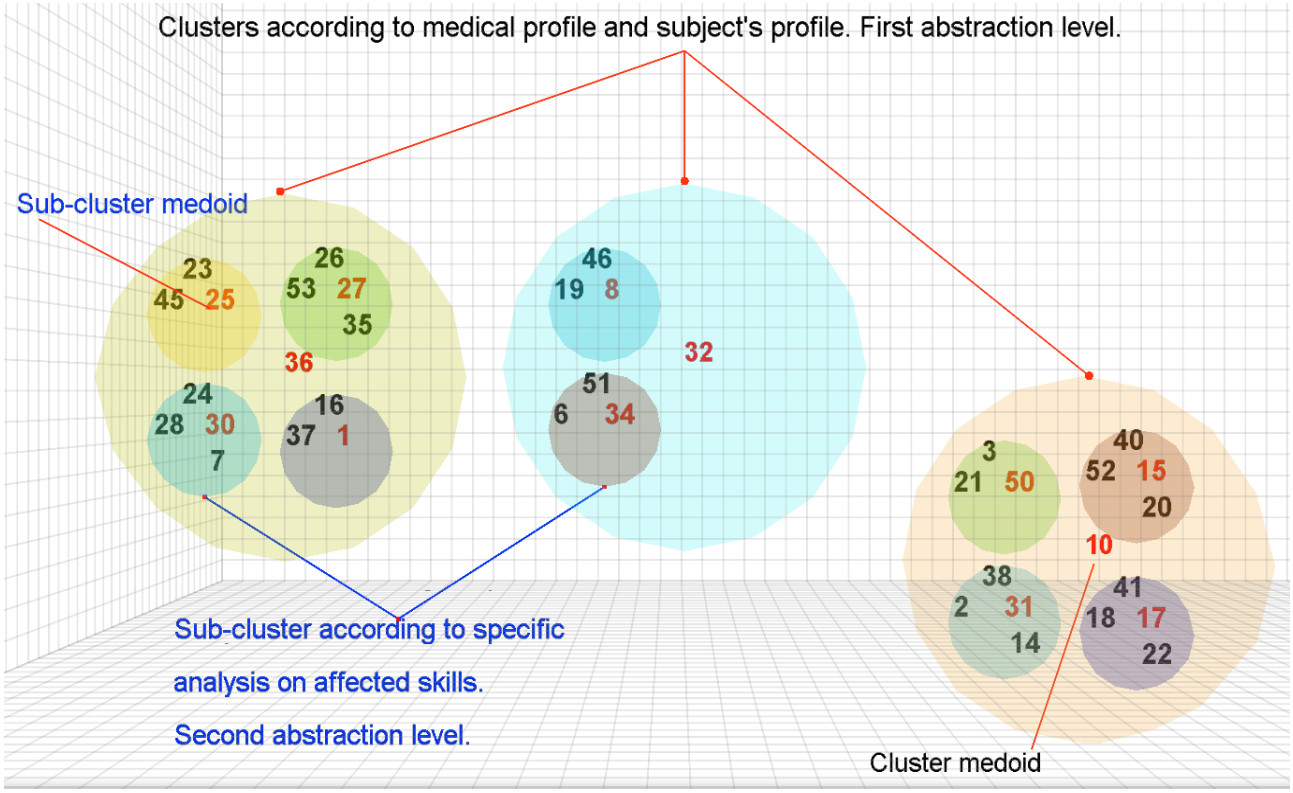
Figure 2. Metric used to determine the distance between two subjects in a specific SL area, according to their profile.

$$d_1(S_i, S_j, A) = \sum_{f \in features_{MR}(A)} ManhDist(f(S_i), f(S_j))$$

Figure 3. Metric used to determine the distance between two subjects within a specific first-level cluster.

$$d_2(S_i, S_j, A) = \sum_{d \in dimensions_{IE}(A)} ManhDist(d(S_i), d(S_j)) \cdot \sum_{f \in features_{CDD}(A)} |f(S_i) - f(S_j)|$$

Figure 4. The clustering approach of the SPELTA system. This structure is used for each speech-language area.



Study Participants and Data Preparation

For the study presented in this paper, the SPELTA expert system was deployed, along with the accompanying tools, in three special education institutions for children in Ecuador: Instituto de Parálisis Cerebral del Azuay (Institute of Cerebral Palsy of Azuay), Fundación "General Dávalos" ("General Dávalos" Foundation), and CEDEI School. Over the course of 2 years (from September 2012 to September 2014), a team of 4 SLPs progressively created a corpus of 117 children profiles, including the corresponding number of therapy plans created manually by themselves and subsequent control evaluations. Some relevant data from the corpus are included in Multimedia Appendix 1. The most common conditions were those of cerebral palsy with/without accompanying dysarthria, dyslalia, epilepsy or dysphasia (n=22), Down syndrome with/without dysarthria or dysphasia (n=19), intellectual disability with/without dysarthria or dysphasia (n=10), autistic disorders (n=9), and fetal alcohol syndrome (n=5). These are the disorders with greatest prevalence in the Ecuadorian province of Azuay.

The corpus is admittedly small and sparse, implying that certain conditions may occur only a few times and many combinations are not included. However, that sparsity is a representative feature of the SLT area because the range of disabilities and communication disorders is so broad that even if two cases have the same medical diagnosis and similar patient profiles, they can require largely different therapy strategies or the support of different assistive technologies. The SPELTA expert system was precisely designed bearing this problem in mind.

The collaborating SLPs used the interfaces and services provided by SPELTA to perform an initial screening of each patient, followed by a personalized evaluation of the 102 SL parameters, and finally, the manual design of a proper therapy plan. As shown in Figure 5, the tools were available on mobile devices as well as desktop computers (see Figures 6-8). The patients could use smartphones or tablets to engage in interactive exercises to evaluate some speech-language skills or to receive memory, motor, hearing, and visual stimulation. The mobile apps proved very useful for SLPs to annotate data about patients who suffer from disabilities that affect their motor skills (eg, cerebral palsy, hemiparesis, hemiplegia) because they allow working in a comfortable space for the patient at work or home. In turn, the desktop apps were most useful with patients in a consulting room or in the rehabilitation centers, and to provide remote assistance.

Figure 5. Diagram of the interfaces and services provided by the SPELTA system.

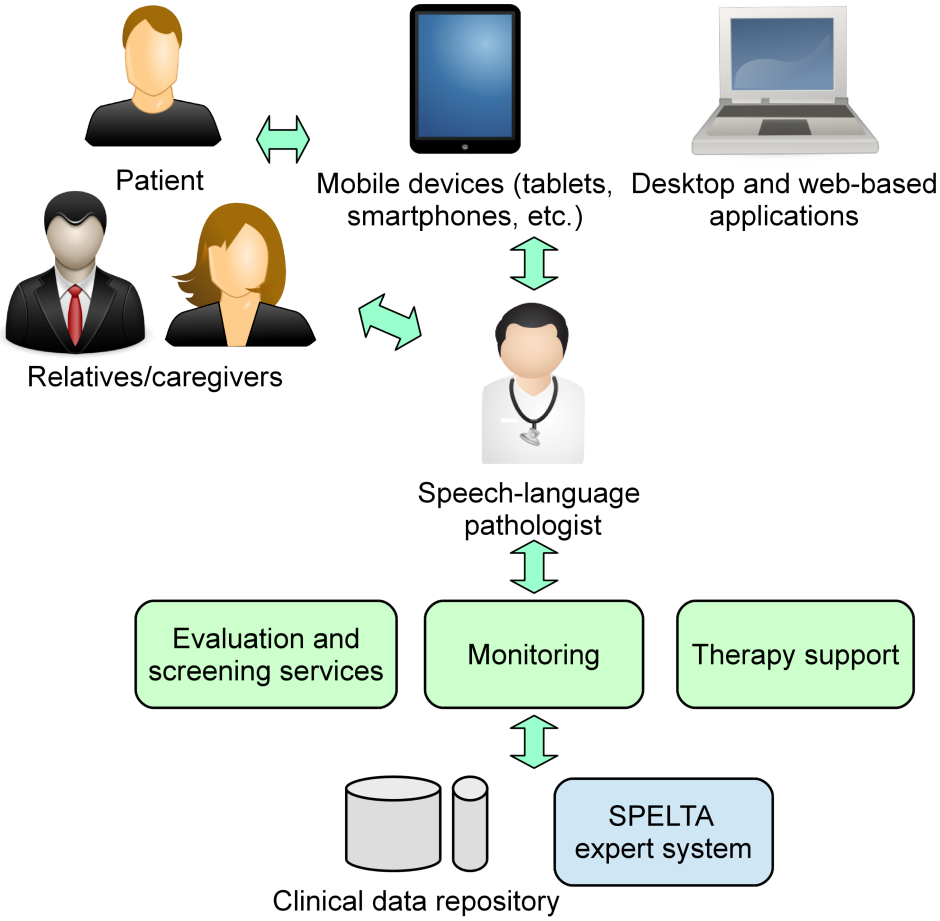


Figure 6. Screen capture of the hearing test that can be applied with mobile devices.

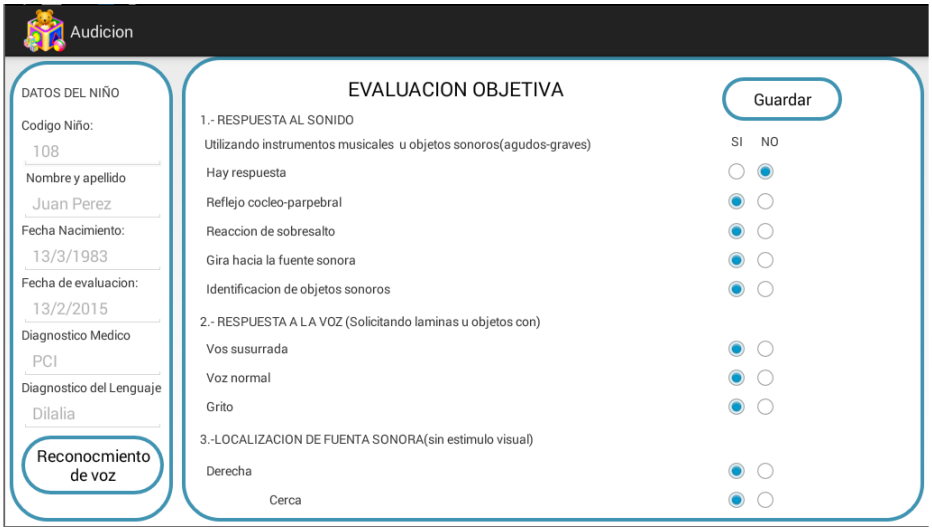


Figure 7. Webpage showing the results of patients' skills in the five SLT areas.

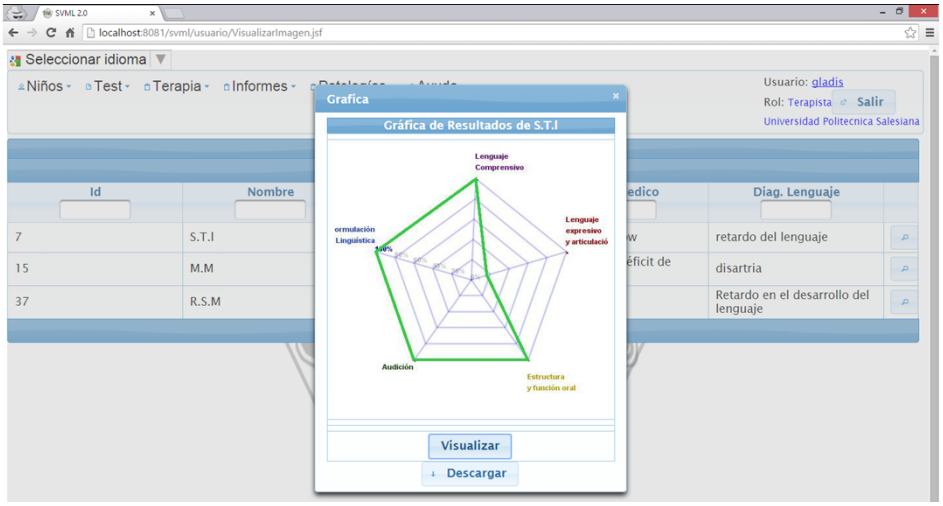
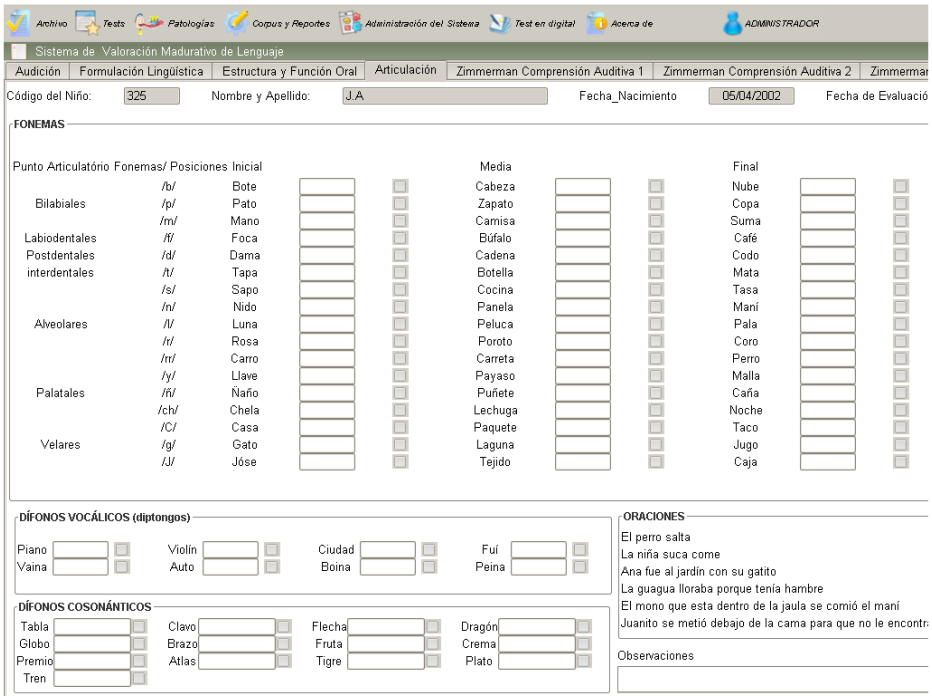


Figure 8. Screen capture of the articulation test on a desktop application.



Evaluation Method

Having trained its algorithms on the corpus of 117 cases and the corresponding plans, the first stage of the evaluation of the SPELTA expert system involved the generation of therapy plans for the cases of 13 new children (see Multimedia Appendix 2). The SLPs discussed whether each one of the automatically generated plans was convenient or not, considering the following criteria.

Accuracy

The exercises and activities selected by SPELTA must be adequate to support the development and rehabilitation of one or more skills related to speech and language. For example, if a patient needs to improve speech production, it is necessary that they have proper breathing conditions and adequate control of their lips and tongue. The accuracy criterion refers to whether the exercises and activities within a plan match the skills that should be improved in the patient.

Consistency

Each patient's profile has different characteristics, such as medical diagnosis, developmental language age, chronological age, etc. The consistency criterion is used to analyze whether a plan contains exercises and activities that can be carried out in a proper way with each patient, bearing in mind their capacity to understand the requests, the affected skills, the developmental gap, etc. For example, cases 23 and 32 (see Multimedia Appendix 1) represent two patients suffering from Down syndrome who had similar developmental language ages (a difference of only 1 month). However, case 23 presented a developmental gap of 2 years and 1 month, whereas case 32 had a 5-year gap. The consistency criterion provides for dealing with these two cases with different activities and exercises, even

though the profiles are similar in terms of medical diagnosis and developmental age.

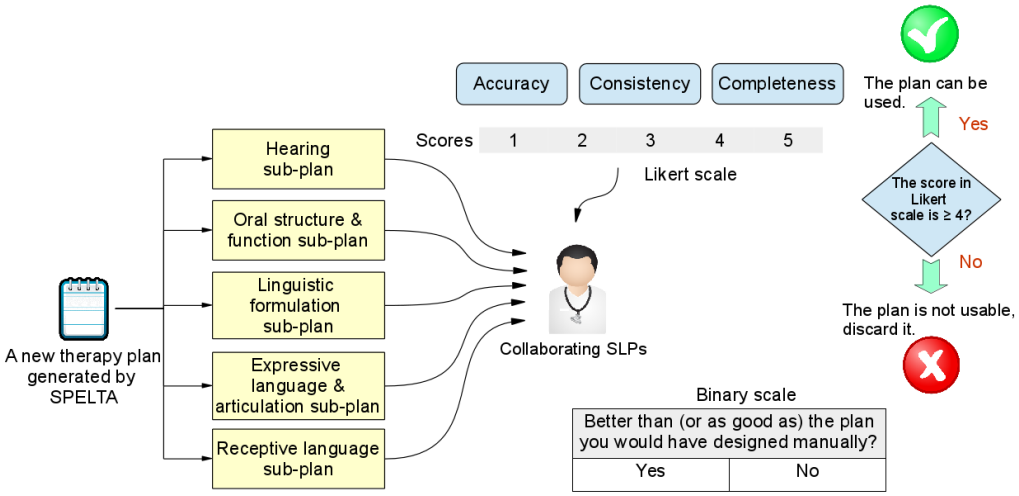
Completeness

In order to have an effective rehabilitation plan, it is necessary to have an adequate number of exercises and activities (not too many or too few). In this line, the completeness criterion is used to determine whether the number and complexity of exercises is adequate for a specific patient. For example, the plan in Table 1 (generated by the SPELTA system) contains the following activities for the hearing area: perform exercises to sounds identification, discriminate sounds of nature, body and animals, and perform phonemes discrimination exercises. The collaborating SLPs confirmed that those guidelines are appropriate to help developing the skills that allow patients to identify phonemes, to construct words and short sentences, and to develop auditory memory over a period of 6 months. Similarly, the number of knowledge areas related to communication is properly delimited for a patient who has a receptive language age of 4 years.

As shown in Figure 9, these criteria were assessed separately for the five subplans of each new plan generated by the SPELTA system, that is, looking at the activities and exercises assigned to each of the five SLT areas. The collaborating SLPs would rate accuracy, consistency, and completeness of each subplan on a 5-point Likert scale, and only the ones that achieved average scores 4 were considered valid and were to be used during the therapy process. Additionally, each SLP would provide a binary response to whether each subplan was “better than” or “as good as” the subplan they would have created manually if given the average time that they could devote to the task.

In order to get further evidence about the statistical significance of the results, we made the experiment to evaluate the SPELTA expert system using a 4-fold cross-validation approach. Specifically, we partitioned the original corpus into 4 sets of 29, 29, 29, and 30 cases, and each cross-validation round consisted of asking the system to provide therapy plans for the cases of each subset, after training it with the cases of the 3 others. The SLPs would discuss whether each one of the automatically generated plans was convenient or not, as above.

Figure 9. The evaluation process followed to assess the plans provided by the SPELTA expert system.



Results

Generation of Therapy Plans for New Cases

Figures 10-14 show the average values obtained on the Likert scale for each of the subplans provided by the SPELTA system when given the input of the 13 new cases: Figure 10 shows the results in the SLT area of hearing, Figure 11 shows oral structure and function, Figure 12 shows linguistic formulation, Figure 13 shows expressive language and articulation, and Figure 14 shows receptive language. The three criteria (accuracy, consistency, and completeness) are represented with different line colors. We can make the following observations per area.

Figure 10. Results achieved by the expert system in the area of hearing.

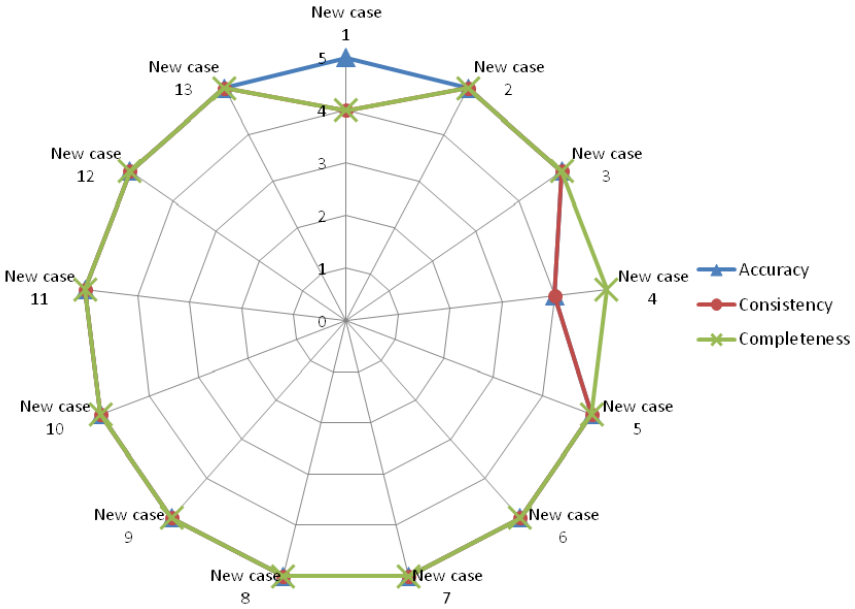


Figure 11. Results achieved by the expert system in the area of oral structure and function.

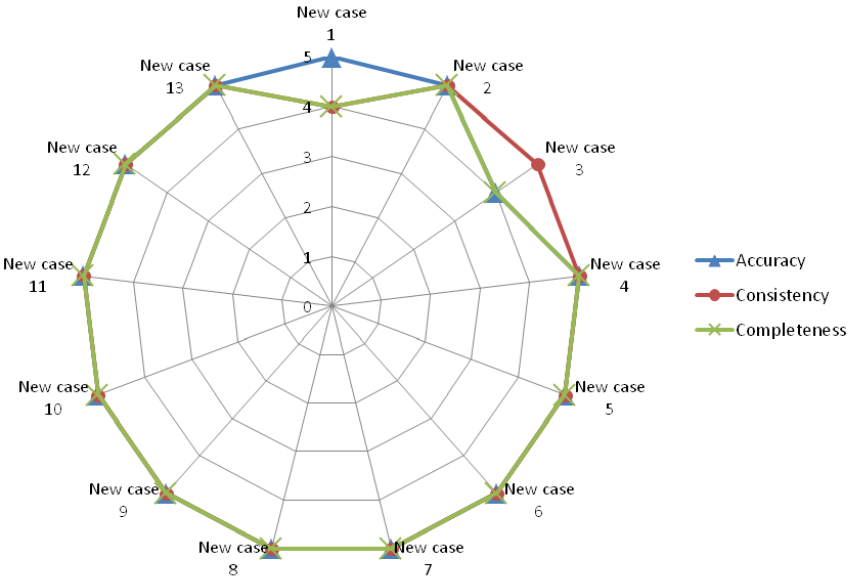


Figure 12. Results achieved by the expert system in the area of linguistic formulation.

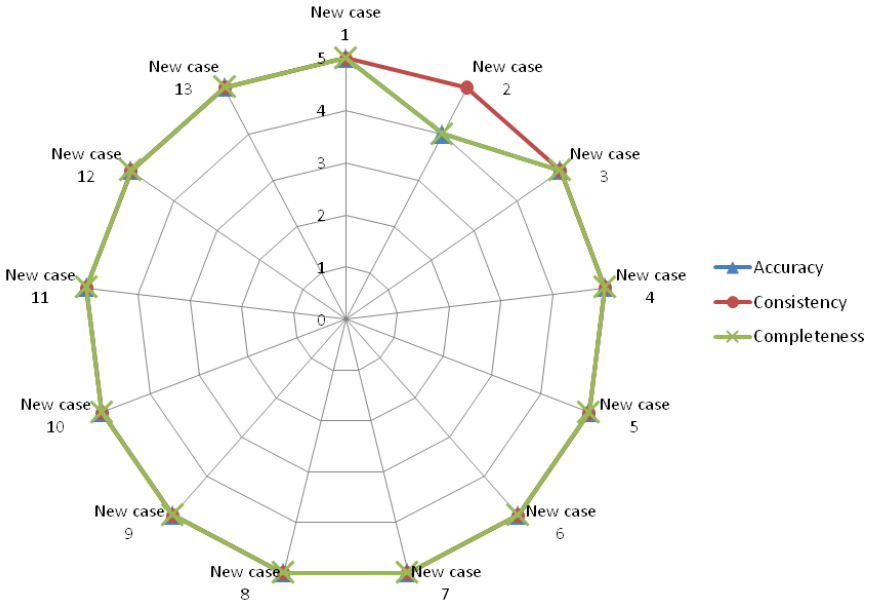


Figure 13. Results achieved by the expert system in the area of expressive language and articulation.

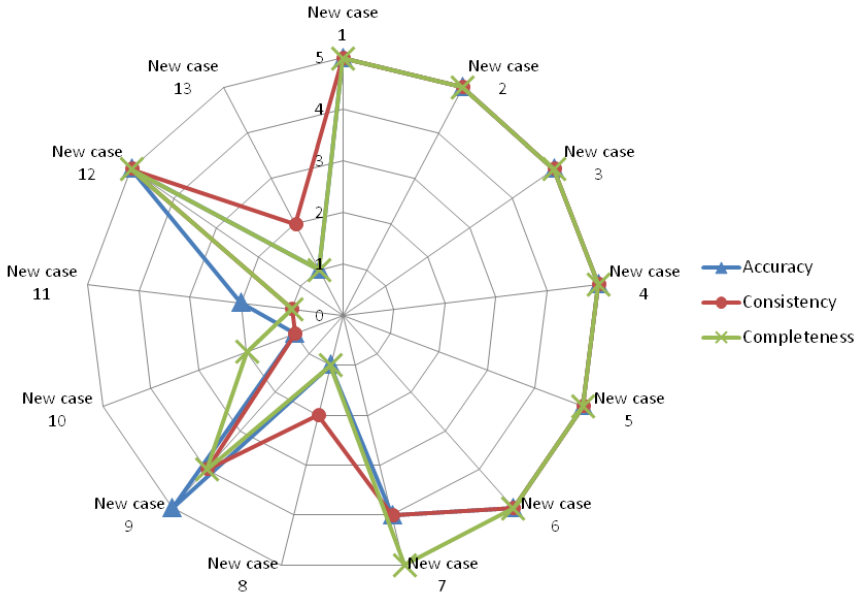
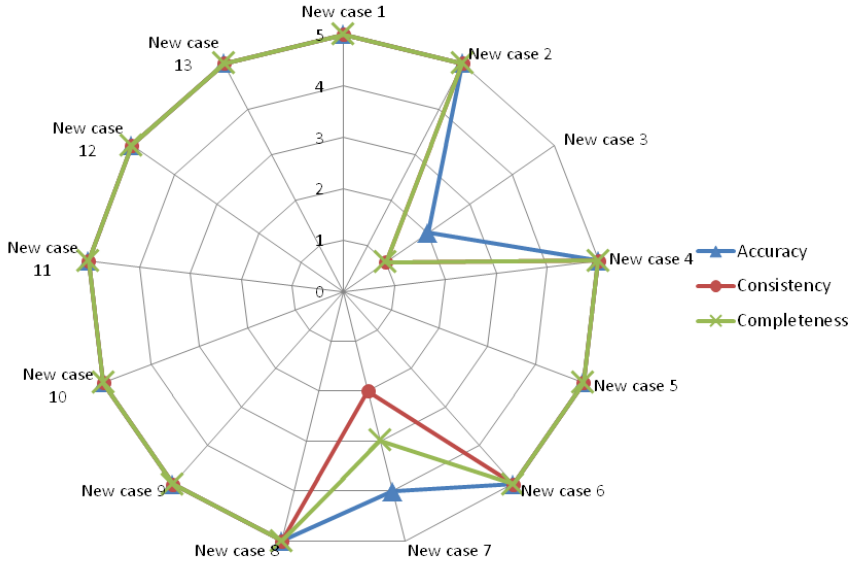


Figure 14. Results achieved by the expert system in the area of receptive language.



Hearing

The 13 subplans generated for this area were considered usable by the SLPs according to the Likert scale. Indeed, only the subplans assembled for new cases 1 and 4 obtained scores of 4 in some of the criteria; all other ratings were 5. For case 1 (a patient with Down syndrome), the SLPs found that it was possible to make some small improvements in the consistency and completeness of the subplan, for which they added one activity to reinforce auditory memory through exercises related to the execution/understanding of simple orders. For case 4 (a patient with mild intellectual disability), in turn, the SLPs determined that the subplan provided by SPELTA was complete for the selected activities, but these did not fully address all the necessary skills in a fully consistent manner for the patient. They changed two activities for less complex ones and added one activity to stimulate the localization of sound sources.

Oral Structure and Function

Again, the 13 subplans generated for this area were considered usable, and only the ones generated for new cases 1 and 3 obtained lower than perfect ratings. Regarding case 1, the SLPs considered the subplan largely usable and were looking at fine-grained details due to their abundant experience in treating Down syndrome. For case 3 (a patient with spastic cerebral palsy and dysphasia), the subplan was found to be fully consistent with the patient's needs, but some of the selected activities were not the best for the case, and the routines missed some exercises the SLPs deemed important. Driven by the most similar cases available in the training corpus, the SPELTA system selected a few exercises that were more suitable for someone with a slightly greater developmental age (around 4 years).

Linguistic Formulation

In this area, all subplans provided by SPELTA were considered usable, and only the one designed for new case 2 (a patient with spastic hemiparesis and dysphasia) got scores of 4 for accuracy and completeness. The SLPs found it necessary to include exercises to complement oral motor rehabilitation and to develop some mainstays (eg, lips control, tongue control) that would provide support in more complex process (eg, getting correct positioning of the phono-articulatory organs for speech production).

Expressive Language and Articulation

This is the area where the expert system showed poorest performance, since it failed to generate usable subplans for new cases 8, 10, 11, and 13. The SLPs found that some of the selected activities would not serve to train the affected skills (inaccuracy), whereas some of the exercises were too complex for the ages and developmental gaps of those patients (inconsistency), and the overall planning of the therapy sessions was not balanced, lacking attention to important traits (incompleteness). The analysis of the cases revealed that the training corpus was too sparse to address their specifics according to the outcomes of the evaluation of the 102 SL parameters. In the absence of very specific training, for example, SPELTA produced largely similar subplans for the new cases 8 and 10, reusing activities and exercises from previous cases that were found to be similar. However, even though both subjects were affected by athetoid cerebral palsy, they differed in that subject 8 would not understand some orders and exercises, whereas subject 10 would not be able to perform some of the selected exercises due to uncontrolled movements of limbs and trunk.

Receptive Language

In this area, the system could not generate a correct subplan only for the cases 3 and 7. The subplan generated for case 7 (a patient affected by cerebral palsy) would have been better suited to someone with greater developmental age, whereas the one generated for case 3 (a patient with spastic cerebral palsy and dysphasia) failed to pay proper attention to the large developmental gap.

The average values of accuracy, consistency, and completeness attained in the five SL areas and globally are shown in Table

2. The validity of the subplans generated automatically and of the therapy plans as a whole (discarding any plan that contained an invalid subplan) are given in Table 3.

Finally, Table 4 summarizes the replies to the question of whether the subplans provided by SPELTA were "better than" or "as good as" the plans that the SLPs would have created manually.

Table 2. Average values of accuracy, consistency, and completeness.

	Hearing	Oral structure & function	Linguistic formulation	Expressive language & articulation	Receptive language	Overall
Accuracy	4.92	4.92	4.92	3.77	4.69	4.65
Consistency	4.85	4.92	5	3.77	4.46	4.60
Completeness	4.92	4.85	4.92	3.77	4.54	4.60

Table 3. Validity of the subplans generated for each area, and the plans as a whole.

Subplans	%
Hearing	100
Oral structure & function	100
Linguistic formulation	100
Expressive language & articulation	69
Receptive language	85
Overall plans for the five areas	54

Table 4. Percentage of positive replies to whether the expert system provided an output comparable to that of a human SLP.

Subplans	%
Hearing	100
Oral structure & function	85
Linguistic formulation	92
Expressive language & articulation	62
Receptive language	85
Overall plans for the five areas	92

Cross-Validation on a Partition of the Corpus

Tables 5,6, and 7 show the average values obtained on the Likert scale in the four rounds of cross-validation with a partition of the original corpus of 117 cases. In turn, Tables 8 and 9 contain data about the validity of the therapy plans and subplans provided by the system, and the replies to the question of whether the subplans were "better than" or "as good as" the plans that the SLPs would have created manually.

Table 5. Average values of accuracy, consistency, and completeness for the areas of hearing and of oral structure and function in the rounds of cross-validation.

K	Hearing			Oral structure & function		
	Accuracy	Consistency	Completeness	Accuracy	Consistency	Completeness
1	4.8	4.74	4.91	4.94	4.97	4.75
2	4.93	4.87	4.9	4.95	4.85	4.87
3	4.84	4.8	4.83	4.82	4.91	4.83
4	4.9	4.72	4.85	4.92	4.83	4.77
Average	4.87	4.78	4.87	4.91	4.89	4.81

Table 6. Average values of accuracy, consistency, and completeness for the areas of linguistic formulation, and of expressive language and articulation in the rounds of cross-validation.

K	Linguistic formulation			Expressive language & articulation		
	Accuracy	Consistency	Completeness	Accuracy	Consistency	Completeness
1	4.84	4.94	4.72	2.93	2.93	2.68
2	4.89	4.98	4.85	2.78	2.91	2.97
3	4.8	4.89	4.91	3.57	3.02	3.31
4	4.9	4.82	4.9	3.14	2.98	3.16
Average	4.86	4.91	4.85	3.11	2.96	3.03

Table 7. Average values of accuracy, consistency, and completeness for receptive language and overall scores in the rounds of cross-validation.

K	Receptive language			Overall		
	Accuracy	Consistency	Completeness	Accuracy	Consistency	Completeness
1	4.57	4.01	4.28	4.42	4.32	4.27
2	4.67	4.41	4.51	4.44	4.40	4.42
3	4.66	4.41	4.55	4.54	4.41	4.49
4	4.34	4.65	4.28	4.44	4.40	4.39
Average	4.56	4.37	4.41	4.46	4.38	4.39

Table 8. Validity of the subplans generated for each area, and the plans as a whole in the rounds of cross-validation.

Subplans	1	2	3	4	Average
Hearing	97%	93%	97%	100%	97%
Oral structure & function	93%	93%	100%	100%	97%
Linguistic formulation	97%	100%	93%	93%	96%
Expressive language & articulation	79%	72%	72%	77%	75%
Receptive language	76%	76%	79%	80%	78%
Overall plans for the five areas	48%	45%	52%	50%	49%

Table 9. Percentage of positive replies to whether the expert system provided an output comparable to that of a human SLP in the rounds of cross-validation.

Subplans	1	2	3	4	Average
Hearing	97%	90%	93%	97%	94%
Oral structure & function	83%	79%	90%	83%	84%
Linguistic formulation	93%	93%	90%	93%	92%
Expressive language & articulation	55%	55%	59%	57%	57%
Receptive language	83%	79%	79%	87%	82%
Overall plans for the five areas	90%	93%	90%	90%	91%

Discussion

Principal Findings

The results obtained in this experiment of generating therapy plans for new subject cases are encouraging about the potential use of the SPELTA expert system in SLT practice. The ratings achieved in terms of accuracy, consistency, and completeness show that the system succeeds in the task of automatically creating new therapy plans out of the knowledge contained in its corpus and in the catalogues of activities and exercises. The subplans generated for the different SL areas were most often considered valid and directly usable, whereas the evaluation of the overall plans was hindered only by the relatively poor performance in the area of expressive language and articulation. Careful analysis of the results in that area suggests that it is necessary to refine some aspects of the reasoning mechanisms of the expert system, even though a more extensive corpus of cases would have also helped to achieve better ratings.

Overall, the SLPs found that the plans provided by SPELTA are, most often, as good as the ones they would have created themselves in their normal work routines (not given sufficient time to work optimally). Thus, the system is a useful tool that can achieve significant savings of valuable and scarce human resources. In order to substantiate the time savings, the SLPs informally measured that the identification and supervision of semi-annual activities to put in a new therapy plan went from an average of 30 minutes down to 5 minutes; the selection of multimedia resources for specific exercises and sessions went from 40 to 6 minutes; and the generation of reports was automated to the point of reducing 24 minutes to 3.

The percentage of positive judgments (92%; Table 4) is much higher than the percentage of plans that contained valid subplans for all five SL areas (54%; Table 3), showing that the SLPs still considered most of the subplans useful and valuable. Accordingly, the SLPs always took the output of SPELTA as a starting point to produce the final therapy plans to use with new patients. Furthermore, they praised the fact that the expert system helped them consider a larger set of activities and exercises than if they had proceeded manually.

The four rounds of the cross-validation experiment yielded similar results, but the fact that the training sets were smaller (87, 88, 88, and 88 cases against 117) had an impact on the quality of the therapy plans, going down from 4.65 accuracy to 4.46, from 4.60 consistency to 4.38, and from 4.60 completeness to 4.39. Still, 49% of the plans were valid straightaway, and 91% were received positively by the SLPs. The greatest impact of working with a more reduced knowledge base was seen in the area of expressive language and articulation, which is in line with the previous observation that a larger corpus will be beneficial.

Comparison With Prior Work

Decision Support Systems (DSS) are becoming increasingly used in the realm of speech and language therapy, with plenty of technical solutions in place to address the specific challenges of the many different disorders. Some DSS depend entirely on input provided by humans, while others rely on signal processing techniques to achieve a level of automation. Thus, on the one hand, Martín Ruiz et al [22] evaluated a Web-based DSS to monitor children's neurodevelopment via the early detection of language delays at a nursery school, relying on input provided by the educators and on a set of over 100 rules to generate alerts in case deviations from the expected developmental milestones. On the other hand, Schipor et al [12] presented a model for automatic assessment of pronunciation quality for children, using Hidden Markov Models (HMM) and implementing a correlation measure to measure the level of intelligibility of utterances. Similarly, Saz et al [13] had used HMM in combination with a subword-based pronunciation verification method. Utianski et al [14] developed an application able to record speech samples and make calculations to assess the integrity of speech production (vowel space area, assessment of an individual's pathology fingerprint, and identification of parameters of the intelligibility disorder). For a final sample, Caballero-Morales and Trujillo-Romero [15] improved the recognition rates for dysarthric patients by integrating multiple pronunciation patterns using genetic algorithms.

All of the aforementioned works focused on providing aids for SLT diagnosis tasks. The idea of aiding in the design of speech and language therapy plans—as we aim to do with the SPELTA system—has fewer precedents in the literature. The closest reference can be found in the work of Schipor et al [16], who developed a system based on fuzzy logic to plan sessions for the treatment of dyslalia, taking input from social, cognitive, and affective parameters, and providing output about types of exercises, frequency, and duration. Later, Yeh et al [17] presented an approach based on neural networks to classify a wide range of SLT problems in order to help design occupational therapy plans, which may include some help to improve communication skills.

Limitations

We believe our study has two main limitations. First, while the results do not show much variability (ratings of 5 were most numerous by far), the SPELTA system needs to be evaluated on a larger set of subject cases. Presumably, the system algorithms will behave more reliably in the presence of a larger corpus, since the sparsity of the corpus we used in our study was one of the reasons for the poor performance in the area of expressive language and articulation.

Second, and probably more important, it would be interesting to experiment with more SLPs from more institutions and other situations than in Ecuador. The 4 SLPs participating in our study had been trained by the same books in the same school, which raises the possibility that there might be some bias in the judgment of the therapy plans presented to them. In the quest for greater evidence, we are actively seeking agreements to test our tools with universities, foundations, and professional associations from other Spanish-speaking countries.

Conclusions

Our study shows that the SPELTA expert system provides valuable input for SLPs to design proper therapy plans for their patients, in a shorter time and considering a larger set of activities than proceeding manually. The system achieves nearly perfect performance in the areas of hearing, oral structure and function, and linguistic formulation, and also decent performance in receptive language. The poorer results in the area of expressive language and articulation have served to identify opportunities for technical improvements, in order to deal properly with new combinations of medical conditions and SL disorders, not properly captured in the corpus. Having a more extensive corpus would obviously help, but in the meantime before a database with thousands of cases becomes available, we are doing research on whether it would be good to adjust internal parameters of the current reasoning system of SPELTA, to define new metrics to compare cases and profiles, and to supplement the internal logic with radically different machine learning artifacts such as the cortical learning algorithm [23].

For future work, we propose a study of two new artificial intelligence techniques supporting the generation of therapy plans. First, we want to use template-based generation methods

with weak supervisions [24], defining a structure based on different levels of granularity in which it will be possible to incorporate common strategies, activities, and resources according to some specific traits and needs derived from the patient's profile. Second, we are interested in deep belief networks and recurrent neural networks [25], which may be able to extract the subtlest patterns from the complex data and interrelations of the SLT area.

Acknowledgments

The authors from Universidad Politécnica Salesiana have been supported by the "Sistemas Inteligentes de Soporte a la Educación" research project (CIDII-010213). We would like to thank Zaituna Bykbaeva, Gladys Ochoa, and all the collaborating people from Instituto de Parálisis Cerebral del Azuay, Fundación "General Dávalos," and CEDEI School.

The authors from the University of Vigo were supported by the European Regional Development Fund (ERDF) and the Galician Regional Government under agreement for funding the Atlantic Research Center for Information and Communication Technologies (AtlantTIC), as well as by the Ministerio de Educación y Ciencia (Gobierno de España) research project TIN2013-42774-R (partly financed with FEDER [Spanish Federation of RD] funds).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Summary of patient profiles.

[PDF File (Adobe PDF File), 67KB - medinform_v4i3e23_app1.pdf]

Multimedia Appendix 2

Summary of profiles for patients randomly selected to the expert system.

[PDF File (Adobe PDF File), 27KB - medinform_v4i3e23_app2.pdf]

References

1. World Health Organization, The World Bank. World report on disability. In: WHO Library Cataloguing-in-Publication Data. Malta: WHO Press; 2011.
2. Ruben RJ. Redefining the survival of the fittest: communication disorders in the 21st century. Laryngoscope 2000 Feb;110(2 Pt 1):241-245. [doi: 10.1097/00005537-200002010-00010] [Medline: 10680923]
3. Lauritsen MB, Pedersen CB, Mortensen PB. The incidence and prevalence of pervasive developmental disorders: a Danish population-based study. Psychol Med 2004 Oct;34(7):1339-1346. [Medline: 15697060]
4. Roy N, Merrill RM, Gray SD, Smith EM. Voice disorders in the general population: prevalence, risk factors, and occupational impact. Laryngoscope 2005 Nov;115(11):1988-1995. [doi: 10.1097/01.mlg.0000179174.32345.41] [Medline: 16319611]
5. Roy N, Stemple J, Merrill RM, Thomas L. Dysphagia in the elderly: preliminary evidence of prevalence, risk factors, and socioemotional effects. Ann Otol Rhinol Laryngol 2007 Nov;116(11):858-865. [Medline: 18074673]
6. American Speech-Language-Hearing Association (ASHA). Speech & Language Disorders. Quick facts URL: http://www.asha.org/About/news/Quick-Facts/ [accessed 2016-02-09] [WebCite Cache ID 6fAxGGqyt]
7. Code C, Petheram B. Delivering for aphasia. Int J Speech Lang Pathol 2011 Feb;13(1):3-10. [doi: 10.3109/17549507.2010.520090] [Medline: 21329405]
8. McAllister L, Wylie K, Davidson B, Marshall J. The World Report on Disability: an impetus to reconceptualize services for people with communication disability. Int J Speech Lang Pathol 2013 Feb;15(1):118-126. [doi: 10.3109/17549507.2012.757804] [Medline: 23323824]
9. Turnbull R, Wehmeyer M, Shogren K. Exceptional Lives: Special Education in Today's Schools (7th Edition). Kansas: Prentice Hall; 2013.
10. Mackenzie C, Le MM, Lendrem W, McGuirk E, Marshall J, Rossiter D. A survey of aphasia services in the United Kingdom. Eur J Disord Commun 1993;28(1):43-61. [Medline: 8400482]
11. Code C, Heron C. Services for aphasia, other acquired adult neurogenic communication and swallowing disorders in the United Kingdom, 2000. Disabil Rehabil 2003 Nov 4;25(21):1231-1237. [doi: 10.1080/0963828031000159961] [Medline: 14578063]
12. Schipor OA, Pentiuc SG, Schipor MD. Automatic Assessment of Pronunciation Quality of Children within Assisted Speech Therapy. Electronics & Electrical Engineering 2012 Jun 11;122(6):15-18. [doi: 10.5755/j01.eee.122.6.1813]
13. Saz O, Yin S, Lleida E, Rose R, Vaquero C, Rodríguez WR. Tools and Technologies for Computer-Aided Speech and Language Therapy. Speech Communication 2009 Oct;51(10):948-967. [doi: 10.1016/j.specom.2009.04.006]

14. Utianski R, Sandoval S, Lehrer N, Berisha V, Liss J. Speech assist: An augmentative tool for practice in speech-language pathology. Journal of the Acoustical Society of America 2013;134(5). [doi: 10.1121/1.4831186]
15. Caballero-Morales S, Trujillo-Romero F. Evolutionary approach for integration of multiple pronunciation patterns for enhancement of dysarthric speech recognition. Expert Systems with Applications 2014 Feb;41(3):841-852. [doi: 10.1016/j.eswa.2013.08.014]
16. Schipor O, Pentiuc S, Schipor M. Improving computer-based speech therapy using a fuzzy expert system. Computing & Informatics 2010;29(2):303-318.
17. Yeh Y, Hou T, Chang W. An intelligent model for the classification of children's occupational therapy problems. Expert Systems with Applications 2012 Apr;39(5):5233-5242. [doi: 10.1016/j.eswa.2011.11.016]
18. Robles-Bykbaev VE, López-Nores M, Pazos-Arias JJ, Arévalo-Lucero D. SPELTA: An expert system to generate therapy plans for speech and language disorders. Expert Systems with Applications 2015 Nov;42(21):7641-7651. [doi: 10.1016/j.eswa.2015.06.011]
19. Robles-Bykbaev V, López-Nores M, Pazos-Arias J, Quisi-Peralta D, García-Duque J. An Ecosystem of Intelligent ICT Tools for Speech-Language Therapy Based on a Formal Knowledge Model. Stud Health Technol Inform 2015;216:50-54. [Medline: 26262008]
20. Geetha T, Arock M. Data clustering using modified k-medoids algorithm. IJMEI 2012;4(2):109-124. [doi: 10.1504/IJMEI.2012.046988]
21. Hariharan M, Chee LS, Ai OC, Yaacob S. Classification of speech dysfluencies using LPC based parameterization techniques. J Med Syst 2012 Jun;36(3):1821-1830. [doi: 10.1007/s10916-010-9641-6] [Medline: 21249515]
22. Martín Ruiz ML, Valero Duboy MA, Torcal LC, Pau de la Cruz I. Evaluating a web-based clinical decision support system for language disorders screening in a nursery school. J Med Internet Res 2014;16(5):e139 [FREE Full text] [doi: 10.2196/jmir.3263] [Medline: 24870413]
23. Numenta INC. Hierarchical Temporal Memory (HTM). 2015. URL: https://github.com/numenta/nupic/wiki/Hierarchical-Temporal-Memory-Theory [accessed 2016-02-09] [WebCite Cache ID 6fAybAP55]
24. Unger C, Bühmann L, Lehmann J, Ngonga NA, Gerber D, Cimiano P. Template-based question answering over RDF data. 2012 Apr 20 Presented at: 21st International Conference on World Wide Web; 2012; Lyon p. 639-648. [doi: 10.1145/2187836.2187923]
25. Deng L. Deep Learning: Methods and Applications. FNT in Signal Processing 2013;7(3-4):197-387. [doi: 10.1561/2000000039]

Abbreviations

DSS: Decision Support Systems
HMM: Hidden Markov Models
ICT: information and communication technology
SL: Speech and Language
SLP: Speech and Language Pathologist
SLT: Speech and Language Therapy
SPELTA: SPEech and Language Therapy Assistant

Edited by G Eysenbach; submitted 18.02.16; peer-reviewed by WY Wang, I Tobolcea; comments to author 31.03.16; revised version received 10.05.16; accepted 11.06.16; published 01.07.16.

Please cite as:
Robles-Bykbaev V, López-Nores M, García-Duque J, Pazos-Arias JJ, Arévalo-Lucero D
Evaluation of an Expert System for the Generation of Speech and Language Therapy Plans
JMIR Med Inform 2016;4(3):e23
URL: http://medinform.jmir.org/2016/3/e23/
doi:10.2196/medinform.5660
PMID:27370070

Original Paper

Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach

Thomas Desautels[1], PhD; Jacob Calvert[1], BS; Jana Hoffman[1], PhD; Melissa Jay[1], BS; Yaniv Kerem[2,3], MD; Lisa Shieh[4], MD, PhD; David Shimabukuro[5], MD; Uli Chettipally[6,7], MPH, MD; Mitchell D Feldman[8], MPhil, MD; Chris Barton[7], MD; David J Wales[9], ScD; Ritankar Das[1], MSc

[1]Dascena, Inc, Hayward, CA, United States

[2]Department of Clinical Informatics, Stanford University School of Medicine, Stanford, CA, United States

[3]Department of Emergency Medicine, Kaiser Permanente Redwood City Medical Center, Redwood City, CA, United States

[4]Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States

[5]Division of Critical Care Medicine, Department of Anesthesia and Perioperative Care, University of California San Francisco, San Francisco, CA, United States

[6]Department of Emergency Medicine, Kaiser Permanente South San Francisco Medical Center, South San Francisco, CA, United States

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

[8]Division of General Internal Medicine, Department of Medicine, University of California San Francisco, San Francisco, CA, United States

[9]Department of Chemistry, University of Cambridge, Cambridge, United Kingdom

Corresponding Author:
Jana Hoffman, PhD
Dascena, Inc
1135 Martin Luther King Drive
Hayward, CA, 94541
United States
Phone: 1 (872) 228 5332
Fax: 1 (872) 228 5332
Email: jana@dascena.com

Abstract

Background: Sepsis is one of the leading causes of mortality in hospitalized patients. Despite this fact, a reliable means of predicting sepsis onset remains elusive. Early and accurate sepsis onset predictions could allow more aggressive and targeted therapy while maintaining antimicrobial stewardship. Existing detection methods suffer from low performance and often require time-consuming laboratory test results.

Objective: To study and validate a sepsis prediction method, *InSight*, for the new Sepsis-3 definitions in retrospective data, make predictions using a minimal set of variables from within the electronic health record data, compare the performance of this approach with existing scoring systems, and investigate the effects of data sparsity on *InSight* performance.

Methods: We apply *InSight*, a machine learning classification system that uses multivariable combinations of easily obtained patient data (vitals, peripheral capillary oxygen saturation, Glasgow Coma Score, and age), to predict sepsis using the retrospective Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)-III dataset, restricted to intensive care unit (ICU) patients aged 15 years or more. Following the Sepsis-3 definitions of the sepsis syndrome, we compare the classification performance of *InSight* versus quick sequential organ failure assessment (qSOFA), modified early warning score (MEWS), systemic inflammatory response syndrome (SIRS), simplified acute physiology score (SAPS) II, and sequential organ failure assessment (SOFA) to determine whether or not patients will become septic at a fixed period of time before onset. We also test the robustness of the *InSight* system to random deletion of individual input observations.

Results: In a test dataset with 11.3% sepsis prevalence, *InSight* produced superior classification performance compared with the alternative scores as measured by area under the receiver operating characteristic curves (AUROC) and area under precision-recall curves (APR). In detection of sepsis onset, *InSight* attains AUROC = 0.880 (SD 0.006) at onset time and APR = 0.595 (SD 0.016), both of which are superior to the performance attained by SIRS (AUROC: 0.609; APR: 0.160), qSOFA (AUROC: 0.772; APR: 0.277), and MEWS (AUROC: 0.803; APR: 0.327) computed concurrently, as well as SAPS II (AUROC: 0.700; APR: 0.225) and SOFA (AUROC: 0.725; APR: 0.284) computed at admission (P<.001 for all comparisons). Similar

results are observed for 1-4 hours preceding sepsis onset. In experiments where approximately 60% of input data are deleted at random, *InSight* attains an AUROC of 0.781 (SD 0.013) and APR of 0.401 (SD 0.015) at sepsis onset time. Even with 60% of data missing, *InSight* remains superior to the corresponding SIRS scores (AUROC and APR, $P<.001$), qSOFA scores ($P=.0095$; $P<.001$) and superior to SOFA and SAPS II computed at admission (AUROC and APR, $P<.001$), where all of these comparison scores (except *InSight*) are computed without data deletion.

Conclusions: Despite using little more than vitals, *InSight* is an effective tool for predicting sepsis onset and performs well even with randomly missing data.

(JMIR Med Inform 2016;4(3):e28) doi:10.2196/medinform.5909

KEYWORDS

sepsis; machine learning; clinical decision support systems; electronic health records; medical informatics

Introduction

Sepsis and its associated syndromes are among the leading causes of worldwide morbidity and mortality [1] and are responsible for placing an enormous cost burden on the health care system [2]. Sepsis, severe sepsis, and septic shock are umbrella terms for a broad and complex variety of disorders characterized by a dysregulated host response to infectious insult. Because of the heterogeneous nature of possible infectious insults and the diversity of host response, these disorders have long been difficult for physicians to recognize and diagnose. A redefinition of sepsis has been recently introduced with the goal of increasing the accurate identification of septic patients in clinical and preclinical settings. This new definition, Sepsis-3 [3], eliminates the traditional ternary classification of sepsis progression from sepsis, through severe sepsis, to septic shock and instead utilizes a two-tier identification system tied to increases in mortality probability. Under the new definition, the term “sepsis” is defined as a “life-threatening organ dysfunction caused by a dysregulated host response to infection [3],” which corresponds most closely with the previously established definition of severe sepsis. Organ dysfunction is defined in practice as an increase in the Sequential Organ Failure Assessment (SOFA) [4] score of at least 2 points. These parameters are associated with in-hospital mortality above 10%. Singer et al [3] define “septic shock” as a classification of sepsis “in which underlying circulatory and cellular metabolism abnormalities are profound enough to substantially increase mortality,” and suggest identifying such patients by a serum lactate measurement above 2 mmol/L and hypotension requiring administration of vasopressors to maintain a mean arterial pressure above 65 mm Hg. Septic shock conditions are associated with in-hospital mortality over 40%. We use this newly proposed definition for sepsis as a gold standard for the implementation of our predictive algorithm, *InSight* [5,6]. *InSight* uses only 8 common measurements (vital signs and other easily assessed bedside measurements, plus age) obtained from electronic health records (EHRs) for the prediction and detection of sepsis in the intensive care unit (ICU) population.

A new bedside scoring system to be used outside the ICU, “qSOFA” (for “quick SOFA”), has been proposed as a screening mechanism to prompt the clinician to further investigate for sepsis or to transfer to a higher level of care [3]. The criteria for qSOFA are at least 2 of the following: respiration above 22/min, altered mentation, or systolic blood pressure below 100 mm Hg. Other scoring systems in current use for the determination or prediction of sepsis include the SOFA score [4], the Modified Early Warning Score (MEWS) [7], the Simplified Acute Physiology Score (SAPS II) [8], and Systemic Inflammatory Response Syndrome (SIRS) criteria [9]. These methods utilize tabulation of various patient vital signs and laboratory results to generate risk scores; however, they do not analyze trends in patient data or correlations between measurements.

The purpose of this study is to validate the *InSight* sepsis prediction method for the new Sepsis-3 definitions using retrospective data consisting of minimal, commonly available EHR variables, and to investigate the effects of data sparsity on its performance. In addition, *InSight* predictive performance will be compared with other existing scores and systems.

Methods

Dataset

This work uses the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)-III version 1.3 dataset [10], compiled from the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA between 2001 and 2012. The MIMIC-III set includes anonymized data from over 52,000 ICU stays and more than 40,000 patients. The *InSight* algorithm uses only the EHR-entered components of the MIMIC-III set, and does not require real-time waveform data or the interpretation of free text notes. The MIMIC-III set includes data logged using the CareVue (Philips) and Metavision (iMDSoft) EHR systems, which handle and store some pieces of information differently. These systems were used at BIDMC from 2001 to 2008 and 2008 to 2012, respectively. Since the original MIMIC-III data collection did not impact patient safety and all data were deidentified in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, the requirement for patient consent was waived by the Institutional Review Boards of BIDMC and the Massachusetts Institute of Technology.

Data Extraction and Imputation

We collect a variety of data from the MIMIC-III dataset to define sepsis onset and calculate the *InSight* score, as well as other scores such as MEWS and SOFA for comparison. All data are extracted from the MIMIC-III set using custom PostgreSQL (PostgreSQL Global Development Group) queries. These measurements are temporally binned using a bin width

of one hour; the measurement values are then averaged within a bin. This process and all subsequent calculations are carried out in MATLAB (The Mathworks, Natick, MA). Missing data are imputed using a "carry-forward" system, where the most recent bin value is carried forward to fill subsequent empty bins. In order to provide a comparison not confounded by different data availability at different times preonset, bins that precede the collection of any measurements of the corresponding type are back-filled with the value of the first subsequent bin with measurements. These processed data are then used in downstream calculations.

Gold Standard

We follow the sepsis definition promulgated by Singer et al [3]. Specifically, Singer et al define sepsis as "life-threatening organ dysfunction caused by a dysregulated host response to infection ... [signified by] an acute change in total SOFA score ≥2 points consequent to the infection." Following the retrospective validation study of Seymour et al [11], we retrospectively equate suspicion of infection with an order for a culture lab draw, together with a dose of antibiotics, within a specified window (see Table 1). Due to limitations of the latest release of MIMIC-III (v1.3), negative cultures (blood and other types) are underreported in the database.

To identify an acute change in SOFA score, we adhere to the definition proposed by Seymour et al. Taking the initial time of the earliest culture draw or antibiotic administration as the time of suspicion of infection, we define a window of up to 48 hours before this time (limited by time of data availability) and 24 hours after this time (limited by time of departure from the ICU). The SOFA score at the beginning of this window is compared with its hourly value throughout this window; if this hourly value is ≥ 2 points higher than the value at the start of the window, we define the first such hour as the onset of sepsis and designate the patient as septic (class 1). If a patient fails to have such an event, we classify them as nonseptic (class 0). If the data required to calculate one of the SOFA subscores is not present in the imputed data, that subscore is given the value 0 (ie, "normal"). We also use a modified version of the SOFA respiration score [12], which avoids requiring information regarding patient mechanical ventilation. Seymour et al were primarily concerned with large-scale identification of septic patients, rather than specifically pinpointing *when* these patients became septic. In contrast, we require this temporal information because we are studying a system that *anticipates* the onset of sepsis.

Table 1. Windows of suspected infection, as defined by the presence of a culture and antibiotic administration, following Seymour et al [11].

First event	Window in which second event must occur
Antibiotics administered	Culture taken in the following 72 hours
Culture taken	Antibiotics administered in the following 24 hours

Selected Clinical Measurements and Patient Inclusion

The learning method employed by *InSight* is flexible with regard to the patient data it uses. For the present work, we have selected systolic blood pressure, pulse pressure, heart rate, respiration rate, temperature, peripheral capillary oxygen saturation (SpO_2), age, and Glasgow Coma Score (GCS). All of these features are nearly universally available at the bedside and do not rely on laboratory tests. There is disagreement about which patient measurements constitute vital signs with the most restrictive definitions only including temperature, heart rate, blood pressure, and respiratory rate, and the most inclusive ones including all of the patient data used in this study with the exception of age [13,14]. Thus, we have collectively labeled the set of measurements used in this study as "extended vitals." Although we train and test our method in the ICU, we note that these or similar features should also be available in other settings. Successful prediction from a minimal set of extended vital signs allows for general application of our approach. This feature is particularly useful for patients that cannot be assessed using other scoring systems (eg, SOFA). We exclude all ICU stays from consideration if any of the following are true: the patient was not at least 15 years old (to eliminate pediatric patients); no measurements were recorded in the ICU; the ICU data was logged using CareVue, rather than Metavision; one or more of the measurements required for our predictor were not recorded at any time during the ICU stay; sepsis onset as defined above occurs, but is more than 500 or less than 7 hours into the ICU stay. The inclusion diagram is presented as Figure 1 and the demographic distribution of patients aged 15 years or more is presented as Table 2. It is important to note that the overall hospital mortality rate of 6.9% for all patients meeting inclusion criteria is significantly lower than the mortality rate for sepsis patients only. This is because the overall study population, as detailed in Table 2, includes patients in all ICU units including low mortality settings like the CSRU. In contrast, the vast majority (over 75%) of infectious disease patients in MIMIC III are in the MICU, which has a median hospital length of stay of 6.4 days and a hospital mortality rate of 14.5% [10].

The requirement that sepsis onset in an included patient occurs be at least 7 hours into their ICU stay is for clarity of presentation. In operation, *InSight* only requires data from the 2 hours preceding prediction time. Given that most patients will have EHR data from a hospital unit that preceded the ICU admission (eg, emergency department, inpatient floor), the predictor will become active at time of admission to the ICU. Notably, the predictor can become active 2 hours after ICU admission at the latest. However, we demonstrate the predictive performance of our approach for various prediction horizons, ie, lengths of time prior to the sepsis onset event. In order for this comparison to not be confounded by differing patient inclusion (varying size and composition) at different horizons, we apply a single, consistent, and conservative inclusion criterion of sepsis onset at least 7 hours into the ICU stay. The requirement that sepsis onset occur within 500 hours (over 20 days) is for convenience of analysis and is minimally restrictive; as shown in Table 2, only 5.1% of patients (1149 patients) have

ICU stays of 12 or more days. Similarly, the requirement that all of the chosen measurements are present during the ICU stay is also for analytical convenience, eliminating less than 500 patients, and need not be strictly applied in practice. We plan to loosen this constraint in future work.

The use of only Metavision patients deserves special discussion. For ICU stays logged using the CareVue system, data about procedures performed (ie, cultures being taken) does not appear in the MIMIC-III database in as detailed and comprehensive a fashion as for ICU stays logged using Metavision. Further, while the MIMIC-III version 1.3 dataset includes information from the BIDMC microbiology lab, reporting positive cultures and the results thereof for all patients, negative cultures are not reported consistently. The combination of these facts means that negative cultures are underreported for CareVue patients. This in turn implies that suspicion of infection, as defined by the cooccurrence of culture and antibiotics, is systematically underrepresented in these ICU stays, resulting in a sepsis prevalence of 3.5% for CareVue patients versus 11.3% for Metavision. In light of this disparity, we chose to exclude CareVue patients from our analyses.

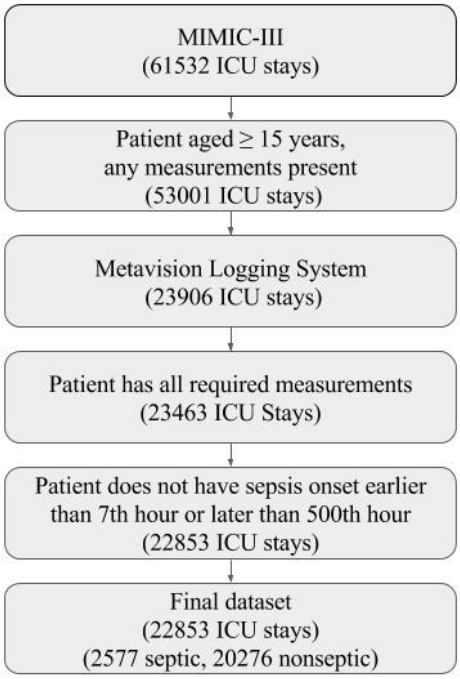
We performed an auxiliary analysis to eliminate patients who received antibiotics prior to the start of their ICU stay (4078 of the 23,906 Metavision ICU stays). This was intended to be a highly sensitive, albeit nonspecific way of removing pre-ICU sepsis cases. Since the exact time-stamp of the start of an ICU stay was not available, we approximated it as 60 minutes prior to initial measurement of any of the extended vital signs from the list in the Clinical Measurements section. Although the 60-minute approximation is discussed here, we also examined various other time windows, and the set of excluded patients was not strongly sensitive to the cutoff time used. With the pre-ICU antibiotic removal, the remaining 19,828 ICU stays were screened identically as previously described, leaving a set of 1840 septic ICU stays and 17,214 nonseptic ICU stays (9.66% sepsis prevalence).

Table 2. Demographics of the included Multiparameter Intelligent Monitoring in Intensive Care version III (MIMIC-III) intensive care unit stays. All stays correspond to patients aged 15 years or more (21,173 hospital admissions).

Demographic characteristic		Number of ICU Stays n (%)
ICU type	medical intensive care unit	9460 (41.89)
	cardiac surgery recovery unit	3345 (14.81)
	surgical intensive care unit	4293 (19.01)
	coronary care unit	2726 (12.07)
	trauma-surgical intensive care unit	2759 (12.22)
Gender	Female	9902 (43.85)
	Male	12,681 (56.15)
Age (years)	15-17	25 (0.1)
Median 65 IQR (53-77)	18-29	982 (4.3)
	30-39	1132 (5.01)
	40-49	2176 (9.64)
	50-59	4038 (17.88)
	60-69	5159 (22.84)
	70+	9071 (40.17)
Length of stay (days)	0-2	15,178 (67.21)
Median 2.0 IQR[a] (1.2-3.8)	3-5	4267 (18.89)
	6-8	1340 (5.93)
	9-11	649 (2.9)
	12+	1149 (5.09)
Death during hospital stay	Yes	1569 (6.95)
	No	21,014 (93.05)

[a]IQR: interquartile range.

Figure 1. Inclusion diagram. All intensive care unit (ICU) stays meeting the sequential inclusion criteria outlined above are included in the training and testing sets. The final dataset has a sepsis prevalence of 11.3%. MIMIC-III: Multiparameter Intelligent Monitoring in Intensive Care version III.



Machine Learning Methods

The training and testing process for the *InSight* prediction system consists of 4 stages: data partitioning, feature construction, classifier training, and classifier testing. The entire training and testing procedure is shown diagrammatically in Figure 2. In the first stage, data are partitioned into 4 folds for cross-validation. Each fold is individually used for testing, while the other 3 folds are concatenated to make the corresponding training set. For each cross-validation fold, feature construction is conducted using the training set. Features include the values of the clinical (vital sign) variables chosen for each of the last 2 hours, denoted x_1 and x_2; continuous, nonlinear function approximations for each posterior probability of sepsis (s=1) given a smoothed estimate of a single clinical variable x_1^i, that is, $P(s=1 \mid x_1^i)$; analogous continuous approximations where $\Delta x^i = (x_1 - x_2)^i$ is the input, $P(s=1 \mid \Delta x^i)$; and tabular approximations to the posterior probability of sepsis, given combinations of discretized versions of 2 or 3 of the Δx^i, that is, $P(s=1 \mid \Delta x^i, \Delta x^j)$ or $P(s=1 \mid \Delta x^i, \Delta x^j, \Delta x^k)$. All of these approximations to posterior probabilities of sepsis are calculated exclusively using the training set. The final feature set is:

$\xi = [x_1, x_2, \ldots P(s=1 \mid x_1^i) \ldots, \ldots P(s=1 \mid \Delta x^i) \ldots, \ldots P(s=1 \mid \Delta x^i, \Delta x^j) \ldots, \ldots P(s=1 \mid \Delta x^i, \Delta x^j, \Delta x^k) \ldots]$

In our first experiment, we assess how performance changes as we use *InSight* to predict whether the patient will become septic at increasingly long times into the future. The *InSight* classifier is given the constructed features and trained to predict whether the patient will be septic (class 1) or not (class 0). This training uses elastic net regularization, which induces a degree of sparsity among the feature weights [15,16]. Finally, the trained classifier is assessed on the disjoint test set; all performance measures presented in this paper are computed on test sets. The entire procedure (fold selection, feature construction, classifier training, and classifier testing) is repeated with independent random partitioning of the data into folds 4 times (ie, 4-fold cross-validation), and for each partitioning, 5 prediction horizons are tested. For each of 0, 1, 2, 3, and 4 hours preceding the time of sepsis onset, we compared *InSight* with qSOFA, MEWS, and SIRS calculated at that time, as well as the SOFA and SAPS II scores computed at ICU admission. While these risk scores are not all sepsis-related, they capture illness severity and represent important benchmarks for performance.

In our second experiment, we test the performance of the *InSight* system in the presence of data sparsity. This situation is simulated by deleting individual EHR-recorded observations according to a random selection procedure. We delete individual observations of the measurements used by our predictor: invasive and noninvasive blood pressure, heart rate, respiration rate, temperature, SpO_2, and GCS. The frequencies with which these values are recorded in the MIMIC-III database are presented in Table 3. These frequencies are on the order of one measurement per hour, close to our temporal discretization frequency. In our experiments, we require that the first measurement of each type for every ICU stay is retained, but all subsequent measurements for every ICU stay may be deleted uniformly at random with a specified probability of deletion, *P*. We set $P = \{0, .1, .2, .4, \text{and } .6\}$ in our experiments. After this random data deletion procedure, we reprocess and impute the data. Note that the gold standard (presence of sepsis and onset time) is determined using the full dataset, and thus is consistent for each ICU stay across all experiments presented here. All subsequent training and testing procedures are similar to the previous experiment.

Table 3. Per-hour observation frequencies among included ICU stays (n=22,853). Three ICU stays were of less than 60 minutes and were discarded from these calculations.

Measurement	Mean (SD) (h^{-1})	Median (IQR[a]) (h^{-1})	Fraction of ICU stays (F[b])
GCS[c]	0.29 (0.16)	0.25 (0.21-0.29)	1
Heart rate	1.31 (3.32)	1.07 (1.01-1.16)	1
Respiration rate	1.30 (3.26)	1.06 (1.00-1.16)	1
SpO_2[d]	1.27 (3.01)	1.06 (0.99-1.17)	1
Temperature	0.31 (0.21)	0.27 (0.23-0.314)	1
NIDiasABP[e]	0.76 (0.39)	0.88 (0.46-1.02)	0.99
NISysABP[f]	0.76 (0.39)	0.88 (0.46-1.02)	0.99
SysABP[g]	0.41 (1.55)	0 (0-0.76)	0.43
DiasABP[h]	0.41 (1.55)	0 (0-0.76)	0.43

[a]IQR: interquartile range.

[b]F: the fraction of these ICU stays with at least one measurement of the given type.

[c]GCS: Glasgow Coma Score.

[d]SpO_2: peripheral capillary oxygen saturation.

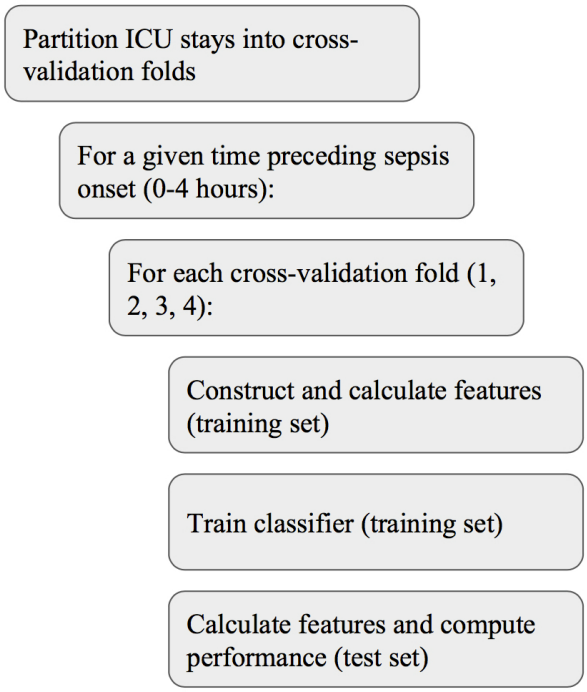
[e]NIDiasABP: noninvasive diastolic arterial blood pressure.

[f]NISysABP: noninvasive systolic arterial blood pressure.

[g]SysABP: invasive systolic arterial blood pressure.

[h]DiasABP: invasive diastolic arterial blood pressure.

Figure 2. Training and testing procedure. The innermost steps in the process (rightmost) are repeated for each partitioning of the data into cross-validation folds (4 partitionings), for each test cross-validation fold in each partition (4 folds), and each time horizon (5 time horizons). ICU: intensive care unit.



Results

The comparison of *InSight* results with each of qSOFA, MEWS, and SIRS, as well as the SOFA and SAPS II scores computed at ICU admission, for sepsis onset and preceding times are presented graphically in Figures 3, 4, and 5. Additional performance measures appear in Table 4. At the time of onset, the *InSight* AUROC (0.8799 [SD 0.0056]) and APR are superior to all of the other methods tested (P<.001 in all cases, assuming normality). This advantage persists at longer preonset prediction times (P<.001 for all AUROC cases and precision-recall for methods other than SOFA; P<.001 and P=.37 for APR against SOFA at 1 and 2 hours before onset, inferior to admission SOFA in APR with P=.001 and P=.009 for 3 and 4 hours before onset).

The ROC curves of *InSight* and the competing scores are shown in Figure 3. As *InSight* is trained to value sensitivity and specificity equally, the ROC curves tend to show a balance between these two constraints. The AUROC advantage held by *InSight* is demonstrated by the form of the ROC curve compared with the other methods (ie, the *InSight* ROC curve generally shows higher sensitivity or specificity, or both, compared with points on the other curves).

Figure 5 shows the area under the precision-recall curves for all scores. precision-recall and ROC curves have a one-to-one correspondence, but emphasize different aspects of the data. While ROC curves are not sensitive to the prevalence of the Class 1 condition (ie, sepsis), the precision value (also known as positive predictive value or PPV) is directly influenced by the prevalence of the Class 1 condition. Further performance measures are presented in Table 4. *InSight* simultaneously achieves moderate sensitivity and specificity, while also attaining good diagnostic odds ratio (DOR) values.

We performed an auxiliary analysis where we eliminated patients who received antibiotics prior to the start of their ICU stay, and the resulting AUROC and model performance metrics were not found to be significantly different from those reported in Figure 3 and Table 4.

We computed the performance of the *InSight* system for random observation deletions, where these occurred with probability $P = \{0, .1, .2, .4, \text{and } .6\}$, with preonset prediction times of 0, 1, 2, and 4 hours. The results of these experiments appear as Figures 6,7, and 8 and Table 5. The typical frequencies of raw data in our patient population (Table 3) are approximately one per hour. Since we discretize time in one-hour intervals, the random data deletions studied here are in a critical regime around the discretization rate and should be expected to affect *InSight's* performance.

Figure 6 shows the ROC curves at selected preonset prediction times and random dropout frequencies. The ROC curves largely maintain performance, even with more than half of all measurements removed. In fact, for predictions 4-hours ahead, and with 60% of measurements missing, *InSight* achieves performance similar to qSOFA detection with no dropout. Full area under ROC and precision-recall curves as a function of time preceding onset are illustrated in Figures 7 and 8, and are further detailed in Table 5.

Table 4. Detailed performance measures for *InSight* and competing scores on the complete Multiparameter Intelligent Monitoring in Intensive Care version III (MIMIC-III) data set, with operating points chosen to make sensitivities close to 0.80. Note that all of quick SOFA's operating points produced sensitivities far from 0.80.

	InSight: 0 hours	*InSight*: 4 hours	SIRS[a]	quick SOFA	MEWS[b]	SAPS II[c]	SOFA[d]
AUROC[e]	0.88 (SD 0.006)	0.74 (SD 0.010)	0.61	0.77	0.80	0.70	0.73
APR[f]	0.60 (SD 0.016)	0.28 (SD 0.013)	0.16	0.28	0.33	0.23	0.28
Sensitivity	0.80	0.80	0.72	0.56	0.70	0.75	0.80
Specificity	0.80	0.54	0.44	0.84	0.77	0.52	0.48
F1[g]	0.47	0.30	0.24	0.39	0.40	0.27	0.27
DOR[h]	15.51	4.75	2.06	6.33	7.85	3.26	3.71
LR+[i]	3.90	1.75	1.30	3.37	3.05	1.57	1.55
LR-[j]	0.25	0.37	0.63	0.53	0.39	0.48	0.42
Accuracy	0.80	0.57	0.47	0.80	0.76	0.55	0.52

[a]SIRS: systemic inflammatory response syndrome

[b]MEWS: Modified Early Warning Score.

[c]SAPS II: Simplified Acute Physiology Score II.

[d]SOFA: Sequential (Sepsis-Related) Organ Failure Assessment.

[e]AURUC: area under the receiver operating characteristic curve.

[f]APR: area under the precision-recall curve.

[g]F1: harmonic mean of precision and recall.

[h]DOR: diagnostic odds ratio.

[i]LR+: positive likelihood ratio.

[j]LR-: negative likelihood ratio.

Table 5. Detailed performance measures of *InSight* when tested and trained with raw data dropouts. Operating points were chosen according to the same procedure as in Table 4.

	InSight, 0 hour, 0% dropout	*InSight*, 0 hour, 10% dropout	*InSight*, 0 hour, 20% dropout	*InSight*, 0 hour, 40% dropout	*InSight*, 0 hour, 60% dropout	*InSight*, 4 hour, 0% dropout	*InSight*, 4 hour, 60% dropout
AUROC[a]	0.89 (SD 0.010)	0.87 (SD 0.006)	0.84 (SD 0.011)	0.83 (SD 0.012)	0.78 (SD 0.013)	0.75 (SD 0.008)	0.73 (SD 0.010)
APR[b]	0.60 (SD 0.022)	0.57 (SD 0.015)	0.54 (SD 0.022)	0.49 (SD 0.021)	0.40 (SD 0.015)	0.27 (SD 0.012)	0.27 (SD 0.009)
Sensitivity	0.80	0.80	0.80	0.80	0.80	0.80	0.80
Specificity	0.82	0.78	0.72	0.68	0.59	0.55	0.52
F1[c]	0.49	0.45	0.40	0.37	0.32	0.30	0.29
DOR[d]	17.90	14.14	10.23	8.31	5.76	4.95	4.38
LR+[e]	4.37	3.62	2.85	2.46	1.95	1.79	1.67
LR-[f]	0.24	0.26	0.28	0.30	0.34	0.36	0.38
Accuracy	0.82	0.78	0.73	0.69	0.61	0.58	0.55

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

[b]APR: area under the precision-recall curve.

[c]F1: harmonic mean of precision and recall.

[d]DOR: diagnostic odds ratio.

[e]LR+: positive likelihood ratio.

[f]LR-: negative likelihood ratio.

Figure 3. Receiver operating characteristic curves for *InSight* versus competing methods at time of onset. MEWS: Modified Early Warning Score; SOFA: Sequential (Sepsis-Related) Organ Failure Assessment; qSOFA: quick SOFA; SAPS II: Simplified Acute Physiology Score II; SIRS: systemic inflammatory response syndrome.

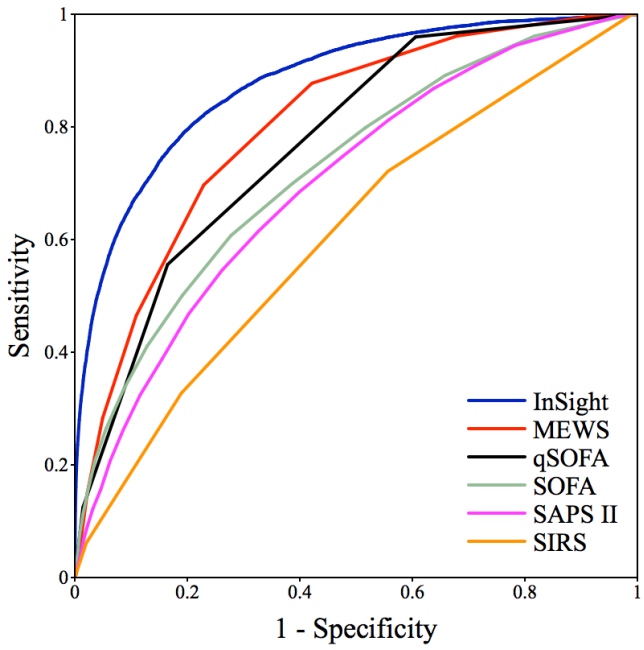


Figure 4. Test set area under receiver operating characteristic curves for *InSight* and competing methods as a function of the amount of time by which prediction precedes potential sepsis onset. Error bars of 1 standard deviation are shown for *InSight*, where the standard deviation is calculated using performance on the cross-validation folds. AUROC: area under the receiver operating characteristic curve; MEWS: Modified Early Warning Score; qSOFA: quick SOFA; SIRS: systemic inflammatory response syndrome.

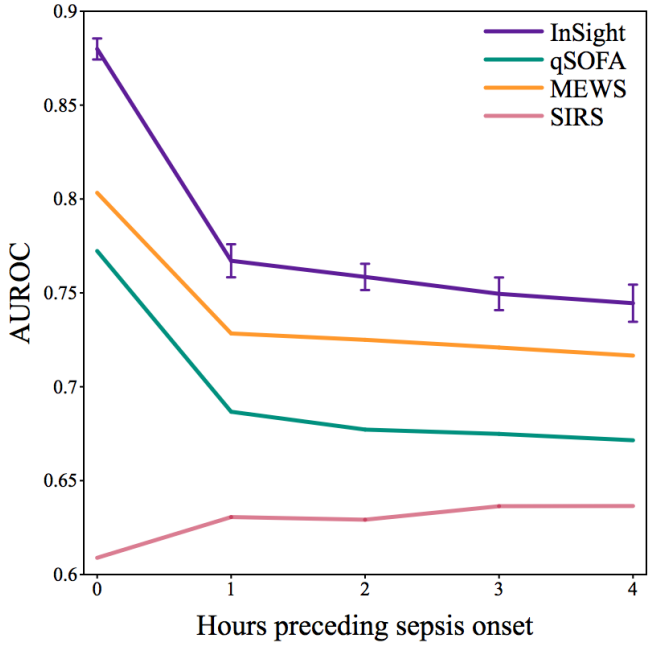


Figure 5. Test set area under precision-recall curves for *InSight* and competing methods as a function of the amount of time by which prediction precedes potential sepsis onset. Error bars of ± 1 standard deviation are shown for *InSight*, where the standard deviation is calculated using performance on the cross-validation folds. APR: area under the precision-recall curve; MEWS: Modified Early Warning Score; qSOFA: quick SOFA; SIRS: systemic inflammatory response syndrome.

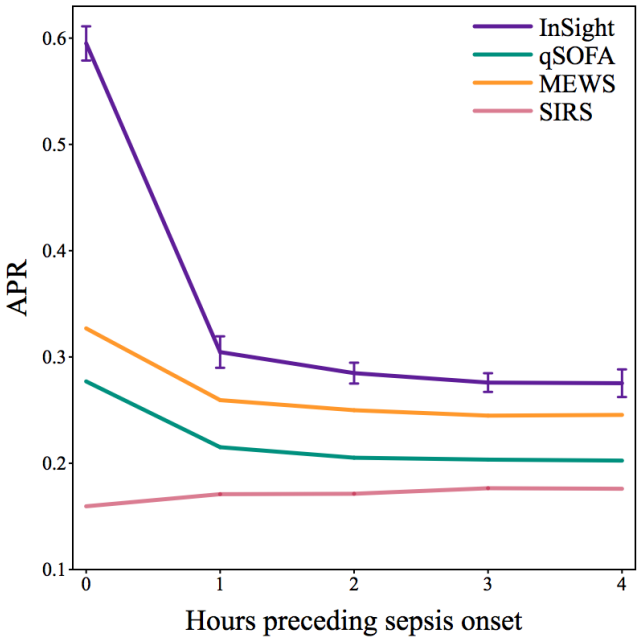


Figure 6. Receiver operating characteristic curves for *InSight* at selected preonset prediction times and random dropout frequencies.

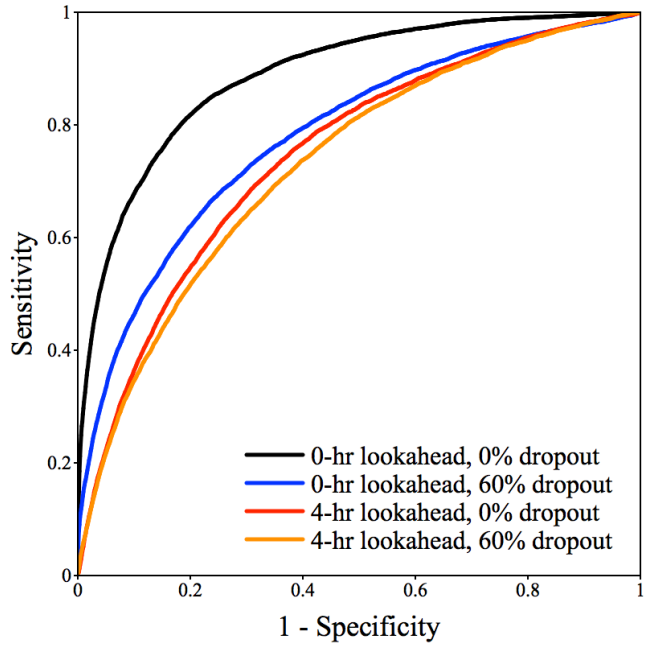


Figure 7. Area under the receiver operating characteristic curve (AUROC) for *InSight* versus preonset prediction time. Each line corresponds to the indicated measurement dropout frequency. All experiments are run with 4-fold cross-validation, with the data repartitioned 4 times.

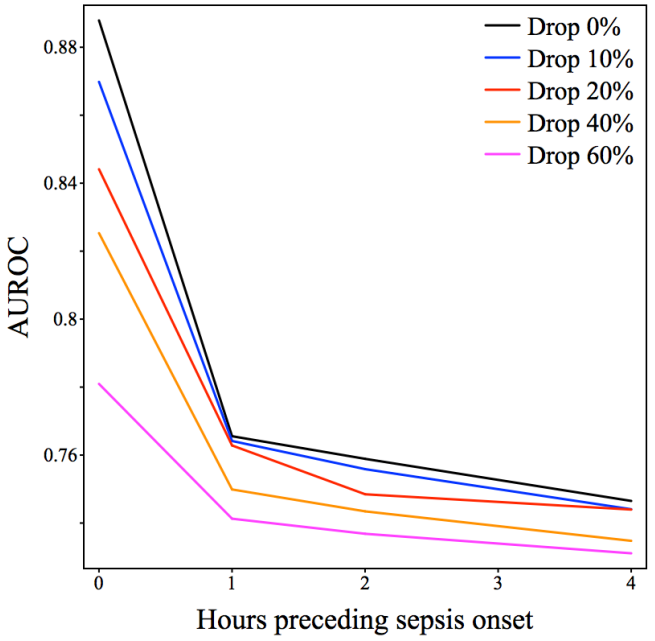
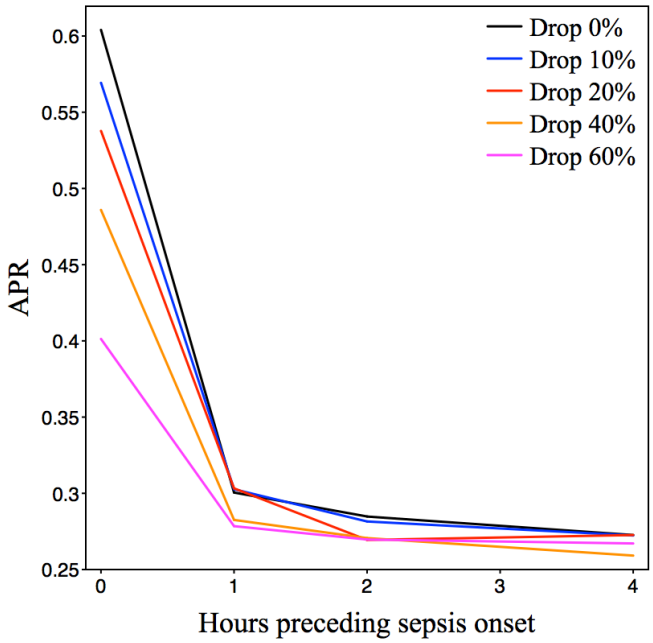


Figure 8. Area under the precision-recall curve (APR) for *InSight* versus preonset prediction time. Each line corresponds to the indicated measurement dropout frequency. All experiments are run with 4-fold cross-validation, with the data repartitioned 4 times.



Discussion

Principal Findings

We tested and validated *InSight*, a machine learning-based system for predicting the onset of sepsis from flexible and minimal data. Using the retrospective MIMIC-III dataset and the new Sepsis-3 definition of sepsis, we trained this system to predict sepsis onset and tested its performance. *InSight* classifies patients (septic vs nonseptic) with a performance that is superior to the corresponding qSOFA, SIRS, and MEWS scores, and it is also superior to the SOFA and SAPS II scores generated at time of admission based on AUROC analysis. It is important to note that MEWS and SAPS II were not explicitly designed for the purpose of sepsis-related severity measurement or prediction. However, these canonical scores represent an important and well-known benchmark for comparison since they are commonly used for sepsis management in clinical settings. *InSight's* superior performance is achieved despite using only age and extended vital sign measurements. All of the extended vital sign measurements (systolic blood pressure, pulse pressure, respiration rate, heart rate, SpO_2, body temperature, and GCS) are commonly available and are easily assessed at the bedside. While the *InSight* system does not offer a manually computable score, it does provide a compelling alternative to the qSOFA and SIRS scores in an increasingly EHR-integrated hospital environment.

Figures 3 and 4 compare the ROC curves of *InSight* with alternative scoring systems. *InSight* generally attains significantly better performance. This result means that, for nearly any specified sensitivity, *InSight* offers superior specificity, and vice versa. Under the gold standard defined above, sepsis has a prevalence of 11.3% (2577/22,853). Furthermore, removing patients who received pre-ICU antibiotics from the analysis did not significantly affect the results. As seen in the precision-recall curve of Figure 5, *InSight's* PPV can easily be operated over 0.5 for 0-hour detection. For prediction one or more hours ahead, a PPV of approximately 0.4 can be obtained if a relatively low sensitivity is acceptable. This would potentially allow narrowly targeted interventions to be applied to a subset of patients whose sepsis diagnosis is nearly certain, while identifying the remaining cases in a more timely manner when their impending sepsis onset becomes more evident.

The detailed numerical results in Table 4 show that *InSight* provides a superior sepsis predictor compared with the alternatives, which tend to have average performance across all measures (SAPS II, MEWS, SOFA) or a large imbalance between sensitivity and specificity (qSOFA, SIRS). While we could choose a different alarm threshold to match or exceed the sensitivity of qSOFA, we would do so at the cost of the other metrics. With respect to the competing scores, the performance of *InSight* stands out, both because it has a high DOR and because it strikes a balance between the other performance metrics without degrading another area. Unlike accuracy, DOR is independent of the prevalence of the positive class. Notably, *InSight* performance 4 hours prior to the onset of sepsis is at least as strong, if not stronger, than the comparison methods.

To improve performance over current scoring systems, *InSight* learns patterns in the trends and correlations among extended vitals through a machine learning process. Several of these extended vitals are also used by SIRS and qSOFA, in conjunction with a suspicion of infection, to diagnose for sepsis, especially outside the ICU setting. The use of correlations in *InSight* is an extension of the approach used by the MEWS scoring system that normalizes patient vitals and sums the results, thereby incorporating some interrelations among different clinical variables. APACHE III also incorporates interrelations among certain variables (eg, pH and pCO_2) via

lookup tables. Similarly, the use of trend information in *InSight* builds on the strategy used by SOFA and APACHE III, where the highest daily value of several patient measurements may be used for score calculations, which implies incorporation of some temporal information.

InSight is also shown by these experiments to be relatively resistant to performance loss from reduced measurement availability. Table 5 presents a variety of performance data for predictors throughout a range of preonset prediction time and random dropout frequency. *InSight* at 40% dropout frequency and at the time of sepsis onset (Table 5) attains performance superior to MEWS at the time of sepsis onset (Table 4). Even with a 60% dropout frequency, *InSight* attains performance that is slightly better than at a prediction time 4 hours before sepsis onset. This result indicates that even if measurement frequency is reduced to well below the prevailing temporal discretization frequency, prognostication is a more difficult task than dealing with measurement dropout. Figure 6, which shows individual ROC curves, and Figures 7 and 8, which show trends across the regime and inter-fold variability, also support this conclusion.

These experiments show *InSight* to be an effective, high performance predictor that uses readily available bedside data for its calculation. This performance is achieved by applying machine learning methods to the relatively simple vital signs data. As noted in the methods section, *InSight* only uses data that would be readily available via ubiquitous monitoring devices (pulse oximeter, blood pressure monitor, etc) and a simple exam. This is a significant difference when compared with the MEWS, SOFA, and SAPS II scoring systems. Additionally, because *InSight* is a machine learning algorithm, it is not restrained to these particular input measurements. In implementation, *InSight* can be trained on the data available in any given setting and will utilize the available measurements that are most relevant to the desired prediction outcome. Of course, performance metrics would be expected to vary with the type and amount of input data available, and training and validation would be required on any novel dataset.

While this is a retrospective study, we are planning future prospective studies through EHR integration of the *InSight* algorithm in an ICU setting. Within that setting, *InSight* has the potential to identify patients at risk of developing sepsis prior to serious patient deterioration or multiple organ failure. *InSight*'s predictive discrimination at 4 hours preceding sepsis onset, as demonstrated in this work, may afford a valuable time window for course-altering clinical intervention. Furthermore, the improvement of sensitivity and specificity over existing sepsis detection methods increases confidence in the accuracy of the *InSight* sepsis alert and therefore may reduce the "alarm fatigue" associated with inaccurate warning systems [17]. Alarm fatigue is defined as the scenario in which too many alarms lead to a decrease in clinician response speed or rate. With increased accuracy and advance warning of impending sepsis, *InSight* has the potential to improve monitoring and treatment for patients who are at risk of sepsis development and to reduce the associated high rates of morbidity and mortality.

Many scoring systems are used for predicting patient outcomes or treatment guidance, despite not being developed for these purposes (eg, SOFA). We present a purpose-built alternative to these systems, based on ubiquitously available vital sign data, for predicting sepsis onset in ICU patients. In this study, *InSight* outperforms all of the other sepsis scoring systems during testing in a variety of realistic conditions. Compared with previous machine learning systems, *InSight* attains similar [18] or better [19,20] AUROC performance at sepsis detection (0.8799 [SD 0.0056], at 0-hours preonset) and offers some prognostic ability while using a significantly more limited collection of patient data [21].

Limitations

There are several practical limitations in this study. First, it is not designed to "discover" a set of rules that could create a manual scoring system. *InSight* is designed as an automatic, EHR-integrated system. Due to its several sequential calculations, including mapping of the input data to a higher-dimensional feature space, *InSight* scores are infeasible to calculate by hand. These calculations are trivial for a computer, however, and can be executed in fractions of a second. Future work may investigate how the *InSight* system can provide clear explanations of its predictions to clinicians including formulae for approximate manual calculations. The gold standard that is based on the Sepsis-3 definitions [3] also presents several difficulties. Sepsis onset is a poorly defined event and identification of an onset time was not the intention of Singer et al; therefore, using their definition for this purpose may be problematic.

We have also chosen to use only a subset of patients in the MIMIC-III (v1.3) database. Because the currently available version of MIMIC-III under-reports cultures, particularly for patients recorded using the CareVue system, we have chosen to work only with patients recorded using the alternative Metavision system to get a more complete picture of suspected infection at various sites. Future work will address these limitations.

An additional limitation is that this study was performed exclusively on ICU data and at a single center, which may limit generalization of our results to other hospitals and hospital systems. While *InSight* operates using only data that are commonly available in nonICU wards, the outcomes reported in this particular study on ICU data do not provide a guarantee of equivalent performance in other settings.

Conclusion

Sepsis prediction is a challenging problem and remains so despite many years of research and development efforts because its manifestation is often unclear until later stages. *InSight* is a machine learning approach specifically designed for this challenge. In this study, *InSight* is shown to be an effective predictor that uses simple and readily available patient data for its calculation. However, in our experiments, the performance of *InSight* is better than the complex, laboratory-value-dependent SAPS II and SOFA scores when computed at ICU admission, and it performs comparably with other machine learning methods in the literature without requiring the laboratory tests

that they incorporate. These experiments also show that *InSight* is resistant to performance degradation from significant random data deletion used to simulate real-world data unavailability. *InSight* is also superior in performance to the qSOFA and SIRS scoring systems that use similar data for calculation. While these two scores have the advantage of being easily computable without computer assistance, *InSight* is readily applicable autonomously in an EHR-integrated environment and offers a high-performance alternative without requiring the collection of any additional data.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1549867. The funder had no role in the conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, and approval of the manuscript; and decision to submit the manuscript for publication.

We gratefully acknowledge the assistance of Dr. Angela J. Rogers, Samson Mataraso, Nima Shajarian, Jasmine Jan, Adrian Gunawan, Allen Chen, and Lauren Song in the preparation of this manuscript. We acknowledge Qingqing Mao and Hamid Mohamadlou for significant contributions to the development and application of the machine learning algorithm, *InSight*.

Conflicts of Interest

All authors who have affiliations listed with Dascena (Hayward, CA, USA) are employees of Dascena.

References

1. Fleischmann C, Scherag A, Adhikari NK, Hartog CS, Tsaganos T, Schlattmann P, International Forum of Acute Care Trialists. Assessment of Global Incidence and Mortality of Hospital-treated Sepsis. Current Estimates and Limitations. Am J Respir Crit Care Med 2016 Feb 1;193(3):259-272. [doi: 10.1164/rccm.201504-0781OC] [Medline: 26414292]
2. Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit Care Med 2001 Jul;29(7):1303-1310. [Medline: 11445675]
3. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016 Feb 23;315(8):801-810. [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
4. Vincent JL, Moreno R, Takala J, Willatts S, De MA, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med 1996 Jul;22(7):707-710. [Medline: 8844239]
5. Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, et al. A computational approach to early sepsis detection. Comput Biol Med 2016 Jul 1;74:69-73. [doi: 10.1016/j.compbiomed.2016.05.003] [Medline: 27208704]
6. Calvert J, Desautels T, Chettipally U, Barton C, Hoffman J, Jay M, et al. High-performance detection and early prediction of septic shock for alcohol-use disorder patients. Ann Med Surg (Lond) 2016 Jun;8:50-55 [FREE Full text] [doi: 10.1016/j.amsu.2016.04.023] [Medline: 27489621]
7. Subbe CP, Slater A, Menon D, Gemmell L. Validation of physiological scoring systems in the accident and emergency department. Emerg Med J 2006 Nov;23(11):841-845 [FREE Full text] [doi: 10.1136/emj.2006.035816] [Medline: 17057134]
8. Le Gall J, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 1993;270(24):2957-2963. [Medline: 8254858]
9. Balk RA. Severe sepsis and septic shock. Definitions, epidemiology, and clinical manifestations. Crit Care Clin 2000 Apr;16(2):179-192. [Medline: 10768078]
10. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35]
11. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, et al. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016 Feb 23;315(8):762-774. [doi: 10.1001/jama.2016.0288] [Medline: 26903335]
12. Jones AE, Trzeciak S, Kline JA. The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation. Crit Care Med 2009 May;37(5):1649-1654 [FREE Full text] [doi: 10.1097/CCM.0b013e31819def97] [Medline: 19325482]
13. Holcomb JB, Salinas J, McManus JM, Miller CC, Cooke WH, Convertino VA. Manual vital signs reliably predict need for life-saving interventions in trauma patients. J Trauma 2005 Oct;59(4):821-8; discussion 828. [Medline: 16374268]
14. The Cleveland Clinic. Vital Signs URL: http://my.clevelandclinic.org/health/diagnostics/hic_Vital_Signs [WebCite Cache ID 6jcU0wOOz]
15. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Statistical Soc B 2005 Apr;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]

16. Calvert JS, Price DA, Barton CW, Chettipally UK, Das R. Discharge recommendation based on a novel technique of homeostatic analysis. J Am Med Inform Assoc 2016 Mar 28 Epub ahead of print(forthcoming). [doi: 10.1093/jamia/ocw014] [Medline: 27026611]
17. Ruskin KJ, Hueske-Kraus D. Alarm fatigue: impacts on patient safety. Curr Opin Anaesthesiol 2015 Dec;28(6):685-690. [doi: 10.1097/ACO.0000000000000260] [Medline: 26539788]
18. Nachimuthu SK, Haug PJ. Early detection of sepsis in the emergency department using Dynamic Bayesian Networks. AMIA Annu Symp Proc 2012;2012:653-662 [FREE Full text] [Medline: 23304338]
19. Stanculescu I, Williams C, Freer Y. Autoregressive hidden Markov models for the early detection of neonatal sepsis. IEEE J Biomed Health Inform 2014 Sep;18(5):1560-1570. [doi: 10.1109/JBHI.2013.2294692] [Medline: 25192568]
20. Stanculescu I, Williams CK, Freer Y. A Hierarchical Switching Linear Dynamical System Applied to the Detection of Sepsis in Neonatal Condition Monitoring. Presented at the Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI). 2014 Jul 23 Presented at: ; 2014; Quebec City, Quebec, Canada p. 752-761.
21. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med 2015 Aug 5;7(299):299ra122. [doi: 10.1126/scitranslmed.aab3719] [Medline: 26246167]

Abbreviations

APR: area under the precision-recall curve
AUROC: area under receiver operating characteristic
BIDMC: Beth Israel Deaconess Medical Center
CCU: coronary care unit
CSRU: cardiac surgery recovery unit
DiasABP: invasive diastolic arterial blood pressure
DOR: diagnostic odds ratio
EHR: electronic health records
F1: harmonic mean of precision and recall
GCS: Glasgow Coma Score
HIPAA: Health Insurance Portability and Accountability Act
ICU: intensive care unit
IQR: interquartile range
LR+: positive likelihood ratio
LR-: negative likelihood ratio
MEWS: Modified Early Warning Score
MICU: medical intensive care unit
MIMIC III: Multiparameter Intelligent Monitoring in Intensive Care version III
NIDiasABP: noninvasive diastolic arterial blood pressure
NISysABP: noninvasive systolic arterial blood pressure
NPV: negative predictive value
PPV: positive predictive value
qSOFA: quickSOFA
ROC: receiver operating characteristic
SAPS II: simplified acute physiology score II
SICU: surgical intensive care Unit
SIRS: systemic inflammatory response syndrome
SOFA: Sequential (Sepsis-Related) Organ Failure Assessment
SpO2: peripheral capillary oxygen saturation
SysABP: invasive systolic arterial blood pressure
TSICU: trauma-surgical intensive care unit

Edited by G Eysenbach; submitted 13.07.16; peer-reviewed by G Bernard, C Coopersmith, JL Vincent; comments to author 04.08.16; revised version received 12.08.16; accepted 29.08.16; published 30.09.16.

Please cite as:
Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, Shimabukuro D, Chettipally U, Feldman MD, Barton C, Wales DJ, Das R
Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach
JMIR Med Inform 2016;4(3):e28
URL: http://medinform.jmir.org/2016/3/e28/
doi:10.2196/medinform.5909
PMID:27694098

Original Paper

Satisfaction Levels and Factors Influencing Satisfaction With Use of a Social App for Neonatal and Pediatric Patient Transfer Information Systems: A Questionnaire Study Among Doctors

Iee Choi[1], MD; Jin Kyu Kim[1], MD, PhD; Sun Jun Kim[1], MD, PhD; Soo Chul Cho[1], MD, PhD; Il Nyeo Kim[1], RN

Research Institute of Clinical Medicine of Chonbuk National University and Biomedical Research Institute of Chonbuk National University Hospital, Department of Pediatrics, Chonbuk National University Hospital, Jeonju, Korea, Republic Of Korea

Corresponding Author:
Jin Kyu Kim, MD, PhD
Research Institute of Clinical Medicine of Chonbuk National University
Biomedical Research Institute of Chonbuk National University Hospital
Department of Pediatrics
Chonbuk National University Hospital
20, Gunjiro, Duckjinku
Jeonju, Korea, 561-712
Republic Of Korea
Phone: 82 63 250 1460
Fax: 82 63 250 1464
Email: kyunim99@gmail.com

Abstract

Background: The treatment of neonatal and pediatric patients is limited to certain medical institutions depending on treatment difficulty. Effective patient transfers are necessary in situations where there are limited medical resources. In South Korea, the government has made a considerable effort to establish patient transfer systems using various means, such as websites, telephone, and so forth. However, in reality, the effort has not yet been effective.

Objective: In this study, we ran a patient transfer information system using a social app for effective patient transfer. We analyzed the results, satisfaction levels, and the factors influencing satisfaction.

Methods: Naver Band is a social app and mobile community application which in Korea is more popular than Facebook. It facilitates group communication. Using Naver Band, two systems were created: one by the Neonatal Intensive Care Unit and the other by the Department of Pediatrics at Chonbuk National University Children's Hospital, South Korea. The information necessary for patient transfers was provided to participating obstetricians (n=51) and pediatricians (n=90). We conducted a survey to evaluate the systems and reviewed the results retrospectively.

Results: The number of patients transferred was reported to increase by 65% (26/40) obstetricians and 40% (23/57) pediatricians. The time taken for transfers was reported to decrease by 72% (29/40) obstetricians and 59% (34/57) pediatricians. Satisfaction was indicated by 83% (33/40) obstetricians and 89% (51/57) pediatricians. Regarding factors influencing satisfaction, the obstetricians reported communication with doctors in charge (P=.03) and time reduction during transfers (P=.02), whereas the pediatricians indicated review of the diagnosis and treatment of transferred patients (P=.01) and the time reduction during transfers (P=.007).

Conclusions: The users were highly satisfied and different users indicated different factors of satisfaction. This finding implies that users' requirements should be accommodated in future developments of patient transfer information systems.

(JMIR Med Inform 2016;4(3):e26) doi:10.2196/medinform.5984

KEYWORDS

social media; personal satisfaction; information systems; patient transfer

Introduction

The treatment of neonatal and pediatric patients in South Korea is limited to certain types of medical institutions depending on disease specificity, patient severity, and treatment difficulty. The amount of medical resources available varies greatly from region to region, with obvious differences in medical infrastructure and administration quality. It is necessary for each region to be equipped with highly trained medical professionals and competent medical facilities, but the availability of such resources is often limited [1-2]. Neonatal and pediatric patients, in particular, are frequently found in emergencies requiring immediate medical attention. When they are transferred from primary or secondary hospitals to tertiary hospitals, a considerable amount of time is often spent locating and identifying available hospital resources, causing significant treatment delays [3-6]. To address this issue, the Emergency Medical Service Act has been enacted in South Korea to strengthen the medical infrastructure. The government has taken the initiative of establishing a National Emergency Medical Center and providing the relevant medical information. Nevertheless, due to information inaccuracy and functional limitations, government authorities and medical professionals have begun to discuss a more efficient emergency medical information system [4].

For the efficient transfer of emergency patients, the emergency medical information provided must be easily accessible and accurate. To this end, social media are perceived as important platforms where users can easily access and share various information. Social media are Web-based services that allow users to form interpersonal networks and to use the networks to connect and communicate with new people [7,8]. The widespread use of mobile phones has enabled real-time communication on social media, and they are currently used in numerous fields due to the efficiency of their information-sharing capabilities. Social media are also widely used in medicine. Notable users include the Centers for Disease Control and Prevention, World Health Organization, and American Public Health Association, which use social media for their information sharing and communication efforts. There are ongoing studies into the use of social media in the medical field and its effectiveness in the United States and other countries [9-13]. In this study, we aimed to develop a model for using the real-time information-sharing function of social media as a patient transfer system. We used a social media platform to create and run a neonatal and pediatric patient transfer information system for obstetric and pediatric physicians in the Jeollabuk-do region of South Korea. We also conducted a questionnaire-based survey to assess the satisfaction with the system and to identify the factors related to satisfaction. We then used the data to identify areas requiring improvement to establish more effective patient transfer information systems in the future.

Methods

Study Design and Participants

The Neonatal Intensive Care Unit and Department of Pediatrics of Chonbuk National University Children's Hospital ran a neonatal and pediatric patient transfer information system (hereinafter, “the Bands”) using Naver Band, which is a closed-type social network service developed by the Internet portal Naver. The Neonatal Intensive Care Unit Band (hereinafter, “NICU Band”) was opened to obstetric physicians since August 2013 and the Department of Pediatrics Band (hereinafter, “DP Band”) was opened to pediatric physicians since November 2014.

The main operators of the NICU Band were the supervising professors and nurse practitioners in the NICU. The nurse practitioners provided daily notifications of the availability of beds and mechanical ventilation equipment, which are essential to patient transfers, so that local obstetricians could take necessary actions based on the information. As most of the neonatal patients transferred to the NICU are in critical condition, transfer notifications of neonatal patients were not usually posted on the Band before the transfer. On the day after the transfer, the professor in charge of the NICU posted a notice of the diagnosis, treatment, and condition of the patient on the Band so that the information was shared with the obstetrician who transferred the patient. The professor also issued daily updates of the condition of any patient who had been transferred to the NICU and was still hospitalized. Training information about neonatal diseases was also provided on the Band, so that the local obstetricians could learn about the diseases and take adequate action when similar situations arise. Information about any potential epidemics was also notified and shared on the Band when they were detected at community care centers or nurseries.

The main operators of the DP Band were the supervising professors and the doctor in charge of the department. Local pediatricians notified the reason for the transfer and condition of the patient on the Band before transferring the patient. A doctor in charge or a professor responded to the local pediatrician about the patient’s condition, diagnosis, and treatment plans in real time. When necessary, a professor or a doctor in charge could also share information about the patient's progress after the diagnosis. In addition, the supervising professors shared information about recent epidemics, the latest treatment guidelines, and any other information that might be useful for the training of local pediatricians in the community. Useful information, about conferences, events, and so on, was also shared on the Band. Pursuant to the Personal Information Protection Act, all personally identifiable information was removed before any information was posted on the NICU Band and DP Band. Our study was approved by the Institutional Review Board of Chonbuk National University Hospital.

Questionnaire

After running the Chonbuk National University Children's Hospital Bands, a survey was conducted with 51 obstetricians and 90 pediatricians who joined the Bands. The professors and doctors in charge, who ran the patient transfer information

system, developed an electronic questionnaire using Google Forms. The questionnaire consisted of 14 questions, spanning 7 pages. Multiple-choice questions were used to query the respondent's department of specialization and sex as well as the duration and frequency of the Band usage. The change in number of patients transferred and time required for transfers was also surveyed to evaluate the effectiveness of the Bands. Satisfaction levels were assessed in 6 categories using Likert scales (5-point scales, with 5 points for very satisfied and 1 point for very dissatisfied) for both categorical satisfaction and overall satisfaction. These 6 categories included information about vacant beds and available equipment in the hospital, information about the transferred patient status, communication with the doctor in charge, rapport with the parents of the patient, decreased time needed for transfer, and checking the diagnosis and confirming the treatment. Short answer questions were used for any additional requests and comments. The survey was tested with the professors and doctors at Chonbuk National University Children's Hospital in the exact same way, as it would be used with local physicians before being conducted with the local physicians. The real closed e-survey was conducted from August 2015 to October 2015. The questionnaire was advertised through the Naver Band and posted on Google Forms. The Web address was sent out to the participants. Only the participants who received the Web address by email and had a Google account could access the site and participate in the survey. The participant entered the site and read the information about the purpose of the survey and how they could participate. Then, they responded to the questions voluntarily. No incentives were offered for participating in the survey. Responses were automatically saved in Google's database. The survey could be submitted once the required questions were answered. Once submitted, the respondent was not allowed to edit or review their responses. The survey was submitted after mandatory questions were answered. Multiple entries from the same individual were not allowed by a built-in function provided by Google Forms. A copy of survey questionnaire can be found in Multimedia Appendix 1.

Data Analysis

The statistical analysis was conducted with SPSS, version 21 (IBM Corporation., Armonk, NY, USA), using frequency analysis and bipartite logistic regression analysis as statistical test methods, with *P* values of less than .05 indicating statistical significance. Frequency analysis was used for the age and sex of the physicians who joined the Bands as well as frequency of usage, number of patients transferred, and time required for transfers to estimate Band usage. For the analysis of factors related to satisfaction, the survey results were divided into a group of highly satisfied respondents (5 points) and a group of all other respondents (4 points and below). Then, binary logistic regression analysis was performed between the 2 groups. We also performed univariate logistic regression analysis and backward multivariate logistic regression analysis to test the correlation between the factors.

Results

Children's Hospital Band Sign-Up and Questionnaire Response

The number of obstetricians in the Jeollabuk-do region who joined the NICU Band was 51 (77% of the total number of obstetricians in the Jeollabuk-do region). Of those, 34 (66%) were male. The number of pediatricians in the Jeollabuk-do region who joined the DP Band was 90 (68% of the total number of pediatricians in the Jeollabuk-do region). Of those, 40 (44%) were male. The questionnaire was answered by 78% (40/51) obstetricians and 63% (57/90) pediatricians. Of the obstetricians who answered the questionnaire, 67% (27/40) were male and 65% (26/40) were aged 40–49 years. Of the pediatricians who answered the questionnaire, 54% (31/57) were male and 51% (29/57) were aged 40–49 years (Table 1).

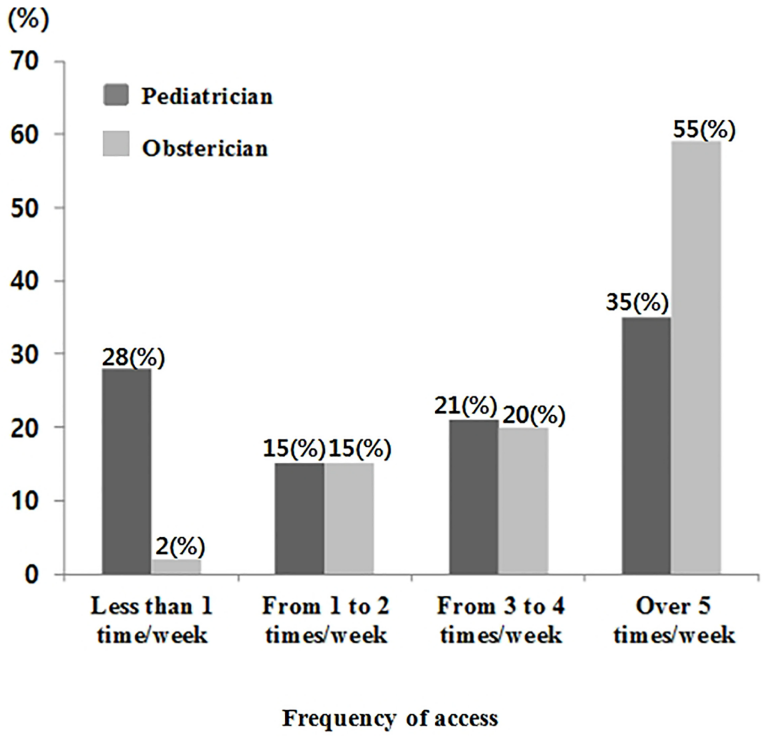
Table 1. Members of the transfer information systems and respondents to the questionnaire.

Factor	Obstetrician in local clinic	Pediatrician in local clinic
Number of physicians who participated in the transfer information system	51	90
Total number of respondents to the questionnaire (%)	40 (78)	57 (63)
Number of male doctors who replied to the survey (%)	27 (67)	31 (54)
Number of members aged between 30 to 39 years (%)	1 (2)	10 (17)
Number of members aged between40 to 49 years (%)	26 (65)	29 (51)
Number of members aged older than 50 years (%)	13 (32)	18 (31)

Frequency and Effects of Using the Children's Hospital Bands

The most common frequency of using the NICU Band, as indicated by 55% (22/40) respondents, was 5 times or more per week. The preferred means of access included using mobile phones by 90% (36/40) respondents and both mobile phones and computers by 10% (4/40) respondents. As for the DP Band, 35% (20/57) respondents used the band 5 times or more per week. The preferred means of access included using mobile phones by 92% (53/57) respondents and both mobile phones and computers by 4% (2/57) respondents (Figure 1). Since using the Children's Hospital Band, 65% (26/40) obstetricians and 40% (23/57) pediatricians responded that the number of patients transferred had increased and 72% (29/40) obstetricians and 59% (34/57) pediatricians responded that the time required for transfers had decreased.

Figure 1. Frequency of access to the patient transfer information systems.



Factors Related to Satisfaction With Using the Children's Hospital Band

In the survey for overall satisfaction with using the Children's Hospital Band, 83% (33/40) obstetricians and 89% (51/57) pediatricians rated it as 4 points or higher (satisfied or very satisfied; Figure 2).

When the factors influencing satisfaction were grouped into 6 categories and the correlation between the factors and satisfaction was tested by univariate regression analysis, the results were statistically significant for both obstetricians and pediatricians (Tables 2 and 3).

Table 2. Univariate logistic regression analysis of each factor that affected satisfaction with the transfer information system by obstetricians in the local clinic.

Factor	Odds ratio	95% CI	*P*
Information about vacant beds and available equipment in the hospital	3.6	1.4-8.9	.005
Information about the status of the transferred patient	12.8	1.5-103.7	.02
Communication with the doctor in charge	6.5	1.9-22.3	.003
Rapport with the parent(s) of the patient	6.4	1.5-27.1	.01
Decreased time needed for transfer	5.1	1.7-14.2	.002
Checking the diagnosis and confirming the treatment	5.2	1.7-15.9	.004

Table 3. Univariate logistic regression analysis of each factor that affected satisfaction with the transfer information system by pediatricians in the local clinic.

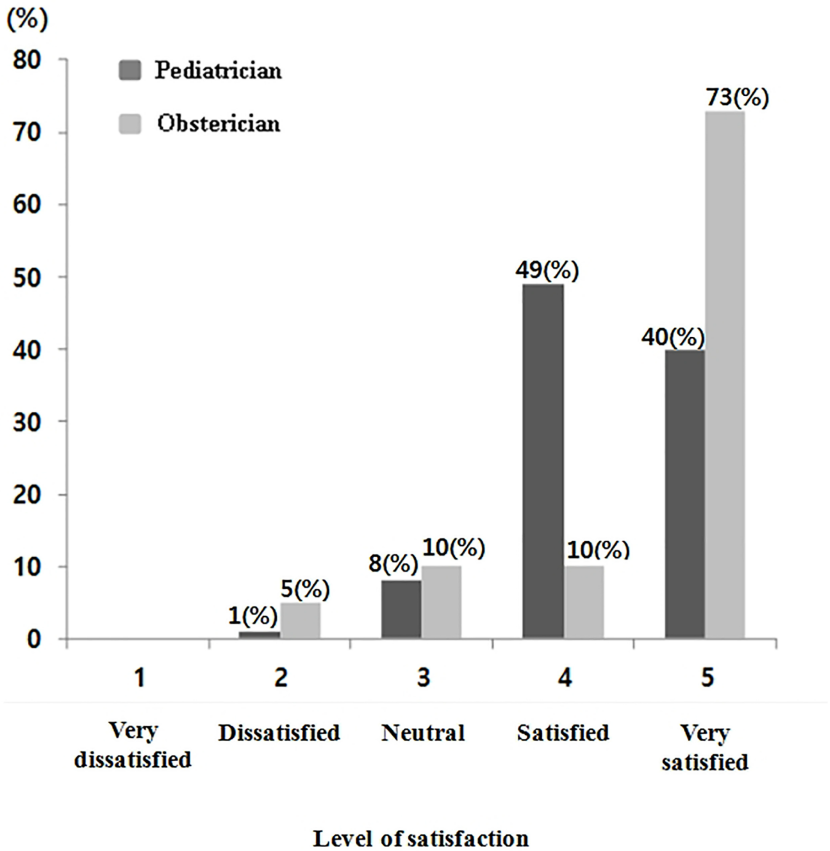
Factor	Odds ratio	95% CI	*P*
Information about vacant beds and available equipment in the hospital	2.5	1.3-4.6	.002
Information about the status of the transferred patient	2.1	1.1-4.3	.02
Communication with the doctor in charge	3.3	1.5-7.2	.002
Rapport with the parent(s) of patient	3.0	1.4-6.5	.004
Decreased time needed for transfer	4.5	1.7-11.7	.002
Checking the diagnosis and confirming the treatment	6.6	2.2-19.7	.001

To identify which of the factors were most strongly correlated with high satisfaction (5 points), backward multivariate bipartite logistic regression analysis was performed between the group with the satisfaction rating of 5 points and the group with the satisfaction rating of 4 points and below, with the data corrected for sex and age. For obstetricians, the ability to communicate with doctors in charge (odds ratio 29, 95% CI 1.311-674.4, *P*=.03) and reduction in time required for transfers (odds ratio 6.5, 95% CI 1.304-37.1, *P*=.02) were highly correlated with satisfaction. For pediatricians, the ability to check the diagnosis and treatment of the patients transferred (odds ratio 3.6, 95% CI 1.276-10.164, *P*=.01) and reduction in time required for transfers (odds ratio 5.6, 95% CI 1.598-19.65, *P*=.07) were highly correlated with satisfaction (Table 4).

Table 4. Adjusted multivariate logistic regression analysis of factors that influence satisfaction with the transfer information systems.

Factor	Odds ratio	95% CI	*P*
Obstetrician of local clinic			
Communication with the doctor in charge	29.7	1.3-674.4	.03
Decreased time needed for transfer	6.5	1.3-37.1	.02
Pediatrician of local clinic			
To check the diagnosis and confirm the treatment	3.6	1.2-10.1	.01
Decreased time needed for transfer	5.6	1.5-19.6	.007

Figure 2. Level of satisfaction with the patient transfer information systems.



Additional Demands and Comments Regarding the Children's Hospital Band

Regarding the additional improvements and developments that the local physicians would like to see in the Children's Hospital Band, 52% (21/40) obstetricians mentioned the need to expand coverage of the Children's Hospital Band to other regions and 25% (10/40) obstetricians mentioned the need for real-time monitoring of hospital beds available for transfers. On the other hand, 49% (28/57) pediatricians mentioned the need for real-time monitoring of hospital beds available for transfers, 29% (17/57) pediatricians mentioned faster responses concerning diagnosis and treatment of transferred patients, and 28% (16/57) pediatricians mentioned concerns over the possible leaking of patient information.

One of the obstetricians expressed the difficulty of patient transfers before using the Children's Hospital Band by saying,

> *Honestly, as an obstetrician, I would do everything to avoid the transfer process altogether. It was*

> *definitely not a pleasant experience getting on the phone and speaking as if I had done something wrong.*

One obstetrician commented on the benefit of using the Children's Hospital Band by saying,

> *Using the Band helps the transfer process a lot. I can now focus on the delivery with greater peace of mind.*

One pediatrician also commented,

> *I am satisfied with the Band because the response I get about the condition of transferred patients is faster than by paper.*

Other comments included,

> *The Band should be expanded to include pediatric surgery, pediatric orthopedics, and many other departments.*

and

> *In addition to the better patient transfer experience, I am also satisfied with other features of the Band, such as information about recent epidemics and refresher training on diseases.*

Discussion

Principal Findings

A social media platform was used to run neonatal and pediatric patient transfer information systems to facilitate communication between Chonbuk National University Children's Hospital and local obstetricians and pediatricians in the Jeollabuk-do region. Analysis of the survey responses from the local physicians showed that the users were highly satisfied. Although each group reported different satisfaction factors and additional demands, both groups saw increased numbers of transferred patients and reductions in time required for the transfers since the transfer system was introduced.

The local physicians were highly satisfied with the Chonbuk National University Children's Hospital Bands as they provided real-time updates on bed information of the regional university hospital and allowed communication about the patients' medical information. Factors associated with satisfaction with the Children's Hospital Bands varied between the obstetricians and pediatricians; the obstetricians' main factor of satisfaction was the ability to communicate with the doctors in charge, whereas the pediatricians regarded the ability to check the diagnosis and treatment information of transferred patients as the most important factor of satisfaction. The difference in satisfaction factors between the obstetricians and pediatricians can be explained in the types of patients transferred. Many of the patients transferred from local obstetric clinics and hospitals are high-risk newborn babies with emergencies occurring immediately after the delivery. In dealing with such patients in need of immediate attention with potentially serious outcomes in survivability and medical disputes, the local physicians regard the ability to identify available hospital beds and communicating with doctors in charge of utmost importance [14]. On the other hand, local pediatric clinics and hospitals tend to transfer patients not for emergency measures but for more advanced diagnosis and treatment [15,16].

Therefore, the varying needs and satisfaction factors of each user group suggest that a customized transfer system is required for each field. Satisfaction with the Children's Hospital Bands can be summarized as the reduction in time required for patient transfers, information sharing, and mutual communication, which is made possible through easy access and provision of information about available hospital beds. Delivering the system on a social media platform can overcome the limitations of existing systems that provide information in one direction only.

Comparison With Prior Work

In previous studies of patient transfer systems, Shin (2007) reported the necessity of establishing region-specific health care systems and transfer systems for South Korea by benchmarking the neonatal patient transfer systems of advanced countries [17], whereas Chang (2011) suggested that the establishment of adequately regionalized patient transfer systems was necessary for efficient neonatal intensive care [1]. Even before the popularization of social media, state-initiated patient transfer systems based on mail, telephone, and the personal computer-era Internet were used in numerous countries, but many were regarded as inefficient. In contrast, the running of our regional patient transfer information system on a social media platform proved to be highly satisfying among local physicians. As suggested in previous studies, various efforts should be made to improve satisfaction with the implementation of transfer systems for the efficient treatment of seriously ill neonatal patients, such as improving accessibility to such transfer information systems and adequately identifying the users' needs, as well as through sufficient leveling of hospitals, regionalization, and the introduction of inter-regional transport systems by reforming the facilities, equipment, and structures.

Limitations

This study has a few limitations. First, the survey was conducted with local physicians who are associated with a single university hospital. Therefore, it would be difficult to generalize the survey results to other regions and other hospitals. In case of Japan, the neonatal patient transfer system assigns a level to each NICU and provides a comprehensive view of hospitals available for transfer [18-20]. However, the Chonbuk National University Children's Hospital Bands were limited to providing information about one hospital only. Efforts could be made to benchmark the Japanese transfer system and group multiple hospitals together in each region for a much more effective system. Second, the Children's Hospital Bands only served the obstetricians and pediatricians in the Jeollabuk-do region who joined the Bands; the survey was conducted with those physicians only. Third, the time required for patient transfers in the survey was based on the physician's perception and not objective measurements. For accurate assessment of actual time reduction, objective remeasurements would be necessary.

Conclusions

In conclusion, the survey of a social media–based patient transfer information system showed that the users were highly satisfied with the provision of information and facilitation of mutual communication, which is necessary for efficient patient transfers. User needs varied depending on the specificity of the

patients transferred. In future developments of patient transfer information systems, the various needs for accessibility to the information system and mutual communication should be accommodated adequately in addition to regionalization and appropriate leveling of hospitals. Furthermore, the patient transfer information system used in our study covered one specific region of a country. A highly effective patient transfer information system would need to go beyond the boundaries of a single region; it needs to connect with other regions and, eventually, connect the whole country. For this reason, we suggest that future studies of patient transfer information systems focus on systems that connect one region to another and on systems that cover the whole nation.

Acknowledgments

This paper was supported by the Fund of the Biomedical Research Institute, Chonbuk National University Hospital (CUH2014-0008). The author thanks all the co-authors for their support in this study. The authors would particularly like to express their appreciation to all local physicians for helping them run the Bands and responding to the questionnaire.

Conflicts of Interest

None declared

Multimedia Appendix 1

Survey questionnaire.

[PPTM File, 106KB - medinform_v4i3e26_app1.pptm]

References

1. Chang YS. Regionalization of neonatal intensive care in Korea. Korean J Pediatr 2011 Dec;54(12):481-488 [FREE Full text] [doi: 10.3345/kjp.2011.54.12.481] [Medline: 22323904]
2. Phibbs CS, Baker LC, Caughey AB, Danielsen B, Schmitt SK, Phibbs RH. Level and volume of neonatal intensive care and mortality in very-low-birth-weight infants. N Engl J Med 2007 May 24;356(21):2165-2175. [doi: 10.1056/NEJMsa065029] [Medline: 17522400]
3. American Academy of Pediatrics Committee on Fetus and Newborn, ACOG Committee on Obstetric Practice. Guidelines for Perinatal Care. 7th edition. Elk Grove Village, IL: American Academy of Pediatrics; 2012.
4. Kim MJ, Lee MC, Yoo J, Kim MJ. Analysis of Maternal and Neonatal Transport by the 1339 Emergency Medical Information Center in Busan Area. J Korean Soc Neonatol 2011;18(1):137-142. [doi: 10.5385/jksn.2011.18.1.137]
5. Ajizian SJ, Nakagawa TA. Interfacility transport of the critically ill pediatric patient. Chest 2007 Oct;132(4):1361-1367. [doi: 10.1378/chest.07-0222] [Medline: 17934123]
6. Stroud MH, Trautman MS, Meyer K, Moss MM, Schwartz HP, Bigham MT, et al. Pediatric and neonatal interfacility transport: results from a national consensus conference. Pediatrics 2013 Aug;132(2):359-366 [FREE Full text] [doi: 10.1542/peds.2013-0529] [Medline: 23821698]
7. Leonardi PM, Huysman M, Steinfield C. Enterprise Social Media: Definition, History, and Prospects for the Study of Social Technologies in Organizations. J Comput-Mediat Comm 2013 Oct 18;19(1):1-19. [doi: 10.1111/jcc4.12029]
8. Boyd D, Ellison N. Social network sites: definition, history, and scholarship. J Comput Mediat Commun 2007;13(1):210-230. [doi: 10.1111/j.1083-6101.2007.00393.x]
9. Chou WS, Hunt YM, Beckjord EB, Moser RP, Hesse BW. Social media use in the United States: implications for health communication. J Med Internet Res 2009;11(4):e48 [FREE Full text] [doi: 10.2196/jmir.1249] [Medline: 19945947]
10. Heldman AB, Schindelar J, Weaver JB. Social media engagement and public health communication: implications for public health organizations being truly “social”. Public Health Reviews 2013;35(1):1-18.
11. Thackeray R, Neiger BL, Smith AK, Van Wagenen SB. Adoption and use of social media among public health departments. BMC Public Health 2012;12:242 [FREE Full text] [doi: 10.1186/1471-2458-12-242] [Medline: 22449137]
12. Park H, Rodgers S, Stemmle J. Health Organizations' Use of Facebook for Health Advertising and Promotion. Journal of Interactive Advertising 2011;12(1):62-77. [doi: 10.1080/15252019.2011.10722191]
13. Avery E, Lariscy R, Amador E, Ickowitz T, Primm C, Taylor A. Diffusion of Social Media Among Public Relations Practitioners in Health Departments Across Various Community Population Sizes. Journal of Public Relations Research 2010 Jul;22(3):336-358. [doi: 10.1080/10627261003614427]
14. Shim JW. The analysis of high risk infant patients being transferred to neonatal intensive care units in Korea. Korea J Perinatol 2012;23(2):87-94.
15. Yoo JW, Lee JH. Clinical analysis of pediatric patients who visited a general hospital emergency center. Korean J Pediatr 2010;53(3):314-322. [doi: 10.3345/kjp.2010.53.3.314]
16. Lee HJ, Park SY, Lee YH, Do BS, Lee SB. Clinical analysis of the pediatric patients seen in the emergency medical center. Korean J Pediatr 2005;48(10):1061-1067.

17. Shin SM. Development of an online system to access the availability of beds and equipment in the referral centers for the transport of newborn patients. J Korean Soc Neonatol 2001;8(1):1-9.
18. Bae C. Perinatal care system for high risk pregnancy and newborn in Japan. Korea J Perinatol 2011;22(4):269-279.
19. Bae CW. Bench-marking of Japanese perinatal center system for improving maternal and neonatal outcome in Korea. Korea J Perinatol 2010;21(2):129-139.
20. Nobuya U. The perinatal care system in Japan. Japan Med Assoc J 2011;54(4):234-240.

Abbreviations

DP Band: Department of Pediatrics Band
NICU Band: Neonatal Intensive Care Unit Band

Edited by G Eysenbach; submitted 17.05.16; peer-reviewed by J Salem, M Anshari; comments to author 10.06.16; revised version received 03.07.16; accepted 20.07.16; published 04.08.16.

Please cite as:
Choi I, Kim JK, Kim SJ, Cho SC, Kim IN
Satisfaction Levels and Factors Influencing Satisfaction With Use of a Social App for Neonatal and Pediatric Patient Transfer Information Systems: A Questionnaire Study Among Doctors
JMIR Med Inform 2016;4(3):e26
URL: http://medinform.jmir.org/2016/3/e26/
doi:10.2196/medinform.5984
PMID:27492978

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org