

Semantic-enhanced models to support timely admission prediction at emergency departments

Jiexun Li · Lifan Guo · Neal Handly ·
Aline A. Mai · David A. Thompson

Received: 3 April 2012/Revised: 19 June 2012/Accepted: 21 June 2012/Published online: 11 July 2012
© Springer-Verlag 2012

Abstract With the rapid outstripping of limited health care resources by the demands on hospital care, it is of critical importance to find more effective and efficient methods of managing care. Our research addresses the problem of emergency department (ED) crowding by building classification models using various types of pre-admission information to help predict the hospital admission of individual patients. We have developed a framework of hospital admission prediction and proposed two novel approaches that capture semantic information in chief complaints to enhance prediction. Our experiments on an ED data set demonstrate that our proposed models outperformed several benchmark methods for admission prediction. These models can potentially be used as decision support tools at hospitals to improve ED throughput rate and enhance patient care.

Keywords Hospital admission prediction · Emergency department · Chief complaint · Classification model · Feature vector · Kernel function

1 Introduction

There is an increasing utilization of emergency department (ED) facilities leading to demand/capacity mismatch. Overcrowding is a growing problem for EDs of many hospitals (Derlet and Richards 2000; Dickinson 1989). It is estimated that one-third of EDs throughout the country experience crowding on a daily basis (Derlet et al. 2001; Wilper et al. 2008). While the number of ED visits is on the rise, patients are facing increasingly long waiting times and ED visit length (Pitts et al. 2008). According to a study of patient visits to 364 US EDs (Hanley and McNeil 1982), in the median ED, a total of 76 % of patients were admitted within 6 h. Only 25 % of EDs admitted more than 90 % of their patients within 4 h and 48 % of EDs admitted more than 90 % of their patients within 6 h. There is a wealth of evidence to indicate the effect of ED crowding and long waiting time on clinical and operational outcomes such as increased ambulance diversions and clinical errors (Bernstein et al. 2009; Miró et al. 1999; Pines et al. 2006; Richardson 2006; Sprivulis et al. 2006). In the United States, patients admitted to the intensive care unit (ICU) suffer higher mortality rates when their ED boarding times exceed 6 h after the decision to admit (Chalfin et al. 2007).

Initially, the problem of ED crowding was thought to be a result of inappropriate use of the ED by patients (increases in numbers of patients arriving) or poor processes within the ED itself. However, according to the studies by the US General Accounting Office and others (Accounting 2003; Steele and Kiss 2008), delays in moving patients who need inpatient care from the ED to inpatient rooms is a more significant contributor to crowding, because the beds these admitted patients occupy in the ED are not available for treating patients still waiting in the waiting room. These delays are typically a result of rooms

J. Li (✉) · L. Guo
College of Information Science and Technology,
Drexel University, Philadelphia, PA 19104, USA
e-mail: jiexun.li@drexel.edu

N. Handly · A. A. Mai
Department of Emergency Medicine, Drexel University
College of Medicine, Philadelphia, PA 19102, USA

D. A. Thompson
Department of Emergency Medicine, Feinberg School
of Medicine, Northwestern University, Chicago, IL 60611, USA

not ready (not cleared or allocated, or even still occupied) and staff not available (whether doctors or nurses).

At present, decisions to admit patients from the ED often are made some hours after arrival because of the need for data gathering, whether by the physician directly or via laboratory and radiological studies. Frequently, the process to prepare the bed and staff is not started until a formal admission decision is made. Streamlining the process to move patients from the ED to an inpatient bed would be enhanced if it were possible to predict those patients who need hospital admission at the moment they arrive in the ED, such as when the first clinical contact occurs at triage and thus beds could be ready when needed. Commonly, data about the patient including a chief complaint, vital signs, past medical history, age, and gender are obtained in triage. This study seeks to use easily available information from triage to predict the need for admission to inpatient care. Such a predictive model can be used at EDs to support admission decisions and start physical preparation of beds and staff allocation in advance. Our ultimate goal is to mitigate ED crowding and improve the quality of patient care.

The remainder of the paper is organized as follows. In Sect. 2, we review related literature in emergency medicine and data mining. In Sect. 3, we introduce a framework of admission prediction using pre-admission information, with specific focus on two semantic-enhanced prediction methods. We evaluate the two proposed methods on a real ED data set in Sect. 4. Finally, we conclude our paper and discuss some limitations and future work in Sect. 5.

2 Literature review

In recent years, a number of data-mining techniques have been applied to health care and prediction models may be the most common applications (Almansoori et al. 2012; Chan et al. 2008). Our study is focused on models for predicting hospital admission. We review related work from two aspects: their use of data and analytical techniques. In addition, we survey previous work on capturing semantic information for data-mining tasks.

2.1 ED data for admission prediction

Prior to the thorough collection of data that occurs once a patient is seen (history, physical examinations, and tests), patient information collected at EDs is limited: demographic characteristics, mode of arrival, reasons for seeking care, brief medical history, and possibly some basic clinical measures (vital signs) are recorded (Chan et al. 2008). These initially gathered data are used to determine whether the patient could be allowed to wait or must immediately

see a doctor. Previous studies predicting admission to inpatient status have used data that might not be typically available at the time of arrival at the ED, such as accommodation type, mobility, lab tests, and even early diagnoses (Chan et al. 2008; Cristianini and Shawe-Taylor 2000; Leegon et al. 2005). These data are only available after the patient has had a sufficient workup, which will usually mean that several hours have passed. Additionally, these data are not always available in ED data sets. Furthermore, pre-admission data can be collected with different methods at EDs and presented in different formats. For instance, gender can be regarded as a Boolean variable and acuity as an ordinal variable, while others such as age and blood pressure can be regarded as continuous variables. Most of these variables are single valued, i.e., each data point (a patient) has only one value for a variable. However, some other variables can be multi-valued. Chief complaints (CCs) are an example.

Chief complaints (CCs) are often considered in prior work to be critical information for decision support at EDs (Chan et al. 2008; Cristianini and Shawe-Taylor 2000; Leegon et al. 2005). A CC is a concise statement describing the patient's reason for seeking medical care, such as symptoms, conditions, and mechanisms of injury. At an ED visit, CCs may be generated by nurses, physicians, and sometimes patients themselves. Examples of CC terms are: fv (fever), nvd (nausea, vomiting, and diarrhea), and sob (shortness of breath). Without a standard lexicon, word variations such as synonyms and acronyms, misspelling, and the institution-specific use of expressions are quite common in CCs, which was often regarded as a major challenge for the use of free-text CCs (Haas et al. 2008; Lu et al. 2008). Moreover, multiple CCs can be assigned to an individual patient to describe his/her different symptoms or conditions, which presents a critical challenge for the representation and modeling of data.

Therefore, raw CCs collected at EDs often require pre-processing to make them useful for further analysis. To deal with misspellings and word variations in CCs, various spell-checking algorithms have been adopted in previous studies (Shapiro 2004). However, most spell-checking algorithms are based on either the edit distance or phonetic similarity; they offer limited value for CC processing. Meanwhile, to deal with major sources of variations in CCs, such as abbreviations, acronyms, and idiosyncratic expressions, it is often required that raw complaints should be converted into standardized terminologies based on a certain coding scheme; such a scheme provides a common standard for CC classification, retrieval, and analysis. Most health care research and systems adopt a general-purpose coding scheme, e.g., the International Classification of Diseases (ICD-9) (ICD-9 2012), the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT)

(Herbert and Hawking 2005), the Unified Medical Language System (UMLS) (Bodenreider 2004), or a domain-specific coding scheme [e.g., reason for visit classification (RVC)]. While general-purpose coding schemes often contain tens of thousands or even millions of terminologies for the standardization of clinical data across multiple sites, domain-specific coding schemes deal with a particular medical domain and therefore contain a smaller number of terminologies (Schneider et al. 1979). Each coding scheme has a hierarchy containing rich semantic relationships among entries, which make them a valuable resource for medical information processing (Achour et al. 2001). Once raw free-text CCs are mapped to standard terms in one of these coding schemes, semantic information embedded in the scheme can be captured to enhance subsequent analysis and prediction. Several studies have proposed approaches to pre-processing and mapping free-text CCs into standardized complaint categories (Chapman et al. 2005; Thompson et al. 2006; Travers and Haas 2004). However, there is not yet a universally accepted approach for CC standardization. Even after the raw CCs are standardized, defining the method to represent and capture the semantic information of coded CCs for decision support remains a challenge.

2.2 Existing techniques for admission prediction

Each type of information collected at EDs tells us certain characteristics about the patients. However, not all of these variables are necessarily contributing factors for predicting hospital admissions. In order to identify good features/variables for prediction, there are two types of feature selection methods: individual feature ranking and feature subset selection (Li et al. 2007). Individual feature ranking is a univariate method that evaluates the relevance of each variable separately based on a certain criteria, such as Pearson correlation coefficient, information gain, and so on. Specifically, in order to identify good features for admission prediction, Pearson's Chi-squared test can be used for categorical variables and Student's *t* test for continuous variables (Chan et al. 2008). Individual feature ranking methods evaluate variables based on their individual predictive power separately without taking into account their interaction effects. On the other hand, feature subset selection methods evaluate the group performance of multiple variables together in a model for prediction.

Many multivariate analytical and machine learning techniques have been used to build prediction models for hospital admission and other similar clinical decisions. Among them, logistic regression is one of the most commonly used techniques because it can take both continuous and categorical variables as independent variables and predict the likelihood of occurrence of an event (Lyon et al.

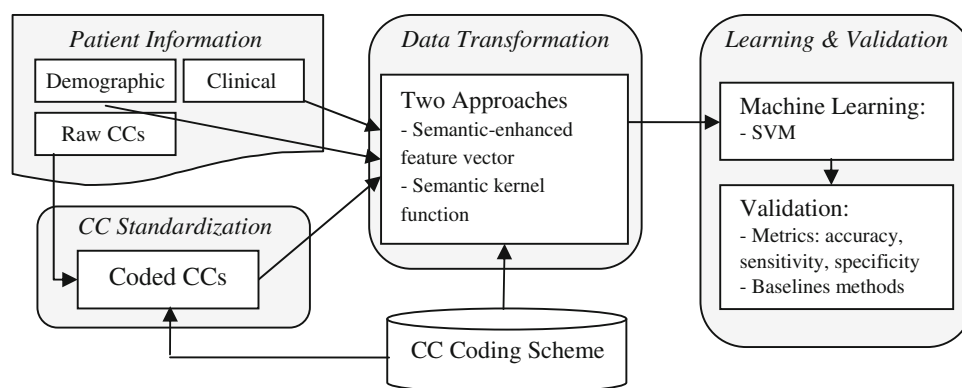
2007). A logistic regression model can also show the relevance of each input variable to the prediction target. In addition, several other supervised learning techniques from the data-mining field have also been applied to clinical decision support, as well. For example, Steele et al. (2006) adopted a decision-tree approach called Classification and Regression Tree (CART) (Lewis and Street 2000) to derive criteria for trauma triage using various field measures (e.g., penetrating mechanism, systolic blood pressure, pulse rate, etc.). Bayesian networks, which can model probabilistic dependencies between input and output variables, have also been used to build prediction models for hospital triage and admission decision support (Leegon et al. 2006; Sadeghi et al. 2006). In addition, advanced computational techniques such as artificial neural network (ANN) can derive complex models for admission prediction with high sensitivity and specificity (Leegon et al. 2006). However, ANN is often criticized for its poor model interpretability.

Despite their underlying computational mechanisms, these prediction techniques often require the representation of each data point, or patient in our case, as a feature vector, where each feature corresponds to an input predictor variable. Categorical variables that can take multiple values, such as chief complaints, can be represented in the model by defining one dummy variable for each value (e.g., each coded CC). However, the hundreds of possible CCs would lead to a feature space of high dimensionality and a very sparse data matrix. Feature selection is often necessary to reduce the dimensionality of feature space for machine learning (Hulse et al. 2012; Li et al. 2007). This problem can be partially solved by treating CCs as one variable and only considering the major complaint for each patient (Leegon et al. 2005). However, this representation is limited because such a nominal CC cannot be multi-valued, i.e., each instance can give only one value for CC. If a patient is assigned with multiple complaints, which is very common, using such a nominal variable cannot sufficiently capture all the complaints but the major complaint alone. Furthermore, the semantic information of the complaints, such as symptoms and systems, cannot be captured by any of these data representations.

2.3 Semantic-enhanced methods for prediction

A feature vector representation is simple and is widely applied in many statistical and machine learning approaches. However, such a representation is not effective in capturing the semantic relatedness between variables. Several measures have been developed to determine the semantic relatedness or similarity between concepts, and they have been shown to enhance the performance of a number of natural language processing (NLP) tasks. For example, semantic similarity or relatedness measures can

Fig. 1 A framework of hospital admission prediction using pre-admission information



be used for word sense disambiguation (WSD) based on the idea that a word should be used in the sense that is most similar to or related to the sense of words that surround it (Leacock and Chodorow 1998; McCarthy et al. 2004; Patwardhan et al. 2003; Rada et al. 1989). Furthermore, through WSD based on word semantic similarity, Ureña-López et al. (2001) showed an improved performance for text categorization (TC). Most of these semantic relatedness measures are defined as a distance between words in certain ontology, which represents knowledge of a set of concepts and their relationships in a domain. WordNet, a freely available lexical database, is such ontology of ~150,000 general concepts as English words. WordNet not only contains definitions of concepts, but also groups them into sets of synonyms called synsets, which compose a hierarchical structure of the various semantic relationships between these synsets (Miller 1990). In addition, other knowledge sources such as Wikipedia also provide a graphical structure among more than three million concepts. Given the high-dimensional space of Wikipedia, Gabrilovich and Markovitch (2007) proposed a method to compute word semantic relatedness by representing the meaning of text as a weighted vector of Wikipedia concepts.

General-purpose knowledge resources such as WordNet and Wikipedia lack domain-specific coverage, which makes them less effective for domain-specific tasks. Given the availability of various ontologies and resources in the biomedical domain, semantic similarity measures can be adapted and applied to medically related tasks. One of the earliest efforts to measure the similarity between biomedical terms was based on Medical Subject Headings (MeSH) as a semantic hierarchy (Miró et al. 1999). Lord et al. (2003) adapted WordNet-based measures to compute the relatedness between terms based on Gene Ontology (GO), a specialized ontology of the molecular functions and biological processes of gene products. The hierarchy of UMLS has also been used as an ontology to compute the path-lengths and “semantic distance” between medical concepts (Bodenreider 2004; Wilbur and Yang 1996).

More recently, Pedersen et al. (2007) derived path-based and content-based measures of semantic similarity and relatedness for biomedical concepts based on the SNOMED-CT ontology and medical corpora. Studies have shown that capturing the semantic information in biomedical terms can improve the performance of tasks such as information retrieval (IR) and classification in the biomedical domain (Li et al. 2007). In our research, we investigated how semantic information of medical terms can be incorporated into a prediction model for enhanced performance.

3 Semantic-enhanced models for admission prediction

In this research, we propose a framework for predicting hospital admission using pre-admission information, as shown in Fig. 1. (1) CC standardization: As one type of patient information collected at EDs, raw chief complaints need to be converted to standard codes based on a certain CC coding scheme. (2) Data transformation: We propose two novel approaches, semantic-enhanced feature vector and semantic kernel function, to transform data into a meaningful format for subsequent model building. (3) Learning and validation: At this step, a model is trained using the processed data for admission prediction and assessed based on standard evaluation metrics. In this section, we will describe each module of the framework in detail.

3.1 Chief complaint standardization

Of the various types of patient information collected, demographic and initial clinical data are well structured and can be easily represented as variables. However, each patient is often assigned several raw (free-text) CCs at EDs, and these CCs are not standardized terms. For example, a symptom of “abdominal pain” may be described as “abdominal pain,” “abdomen pain,” “pain in abdomen,” “ap,” etc. Including all of these raw CCs as variables may

lead to a high-dimensional feature space with an excessive amount of noise. Therefore, we first need to standardize the raw CCs by dealing with problems, such as synonyms, acronyms, misspelling, etc.

In our proposed framework, a CC coding scheme is a critical component in that it is used for both CC standardization and data transformation. The CC standardization process maps each free-text CC to a standard coded CC defined in the coding scheme. In the data transformation process, our algorithms will leverage the structure of CC coding scheme and capture the semantic information of CCs. It is worth noting that any existing CC coding scheme should be applicable in this framework. In our study, we chose to use the Coded Chief Complaints for Emergency Department Systems (CCC-EDS), a comprehensive and granular scheme of chief complaints describing the reason for an ED visit (Thompson et al. 2006). The CCC-EDS was developed based on the study of various ED CC literature and coding schemes, including RVC and ICD-9. The CCC coding engine is a computerized text-parsing algorithm for automatically reading and classifying English language free-text chief complaints into one of these coded chief complaints. The CC algorithm is based on a mapping table that links over 8,000 distinct CC entries into 238 clinically actionable CCs, each corresponding to a specific ICD-9 code.

In CCC-EDS each coded CC C_i is described by its name (w_i), type (t_i), system (s_i), and acuity rating (r_i). There are 8 different type groupings and 14 different system groupings in the scheme (see Table 1). The CCC acuity rating is a five level index for triage assessment based on the Emergency Severity Index (ESI 2012; Wuerz et al. 2000), with a rating of 1 being the most acute and of 5 being the least acute. For example, the chief complaint “chest pain” has type = “illness–symptom,” system = “cardiovascular and immune System,” acuity rating = “2,” and ICD-9 code = 786.50. This coding scheme allows for the subsequent rollup of chief complaint volume statistics into manageable reporting groups.

3.2 Data transformation

CC standardization allows us to significantly reduce the thousands of complaints to a much more manageable size. However, we still need to find a meaningful way of representing coded CCs, demographic information, and other clinical measurements for each patient to train a predictive model. The data transformation module takes different types of patient information as input and transforms it in a certain format that can be input into a machine learning algorithm. Machine learning techniques can be categorized into feature methods and kernel methods. In feature methods, each data object is represented as a feature vector $\mathbf{X} = (x_1, x_2, \dots, x_p)$ in a p -dimensional space. Features need

Table 1 Chief complaint types and systems in CCC-EDS

CC types	CC systems
Injury–trauma	General
Symptom	Eyes
Disease or condition	Ears, nose, mouth, throat, and face
Medical device, procedure, or surgery	Cardiovascular
Exposure	Respiratory
Treatment	Gastrointestinal
Bites or stings	Genitourinary
Other administrative	Musculoskeletal
	Skin, hair, nails, and breast
	Neurological and spine
	Psychiatric and behavioral
	Endocrine
	Hematologic–lymphatic
	Allergic–immunologic

to be predefined to describe certain characteristics or properties of data object. However, if data objects have some complex structure such as a hierarchy or network, features cannot be easily defined to capture the structural information. Kernel methods are an effective alternative to feature methods (Cristianini and Shawe-Taylor 2000). They define a kernel function between a pair of data objects. Formally, a kernel function is a mapping of $K: \mathbf{X} \times \mathbf{X} \rightarrow [0, \infty)$ from an input space \mathbf{X} to a similarity score $K(x,y) = \phi(x) \cdot \phi(y) = \sum_i \phi_i(x) \phi_i(y)$, where $\phi_i(x)$ is a function that maps \mathbf{X} to a higher dimensional space. A kernel function is required to be symmetric and positive semidefinite. Such a kernel function makes it possible to compute the similarity between objects in their original representation without enumerating all of the features. By applying a kernel function to all instance pairs in a training set of size n , we can get an $n \times n$ kernel matrix, which can be fed into a kernel machine, e.g., a support vector machine (SVM) (Cristianini and Shawe-Taylor 2000), to train a model for prediction. The performances of kernel methods are mainly determined by the selection and design of the kernel functions. Another advantage of kernel methods is that they transform heterogeneous data representations into kernel matrices of the same format, which enables the integration of different information (Lanckriet et al. 2004). Kernel-based methods have been frequently used in various machine learning areas, such as pattern recognition (Chapman et al. 2005), data mining (Zhou and Wang 2005), and text mining (Li et al. 2008).

In order to support timely admission prediction at EDs, we propose two “semantic-enhanced” models: a semantic feature (SF) method and a semantic kernel (SK) method. We call them “semantic-enhanced” models because they

both capture the semantic relatedness between complaints by considering their attributes, such as type, system, and acuity rating as defined in the CC coding scheme.

3.2.1 Semantic feature method

First, we consider a feature method and try to represent each patient visit as a feature vector. Each variable of the demographic and clinical information often takes a single value and can be easily represented as a feature. Since multiple chief complaints can be assigned to each patient, each coded CC can be regarded as an individual feature which can take a Boolean value to indicate whether this CC is assigned to the patient or not. However, such a representation often generates a very sparse data matrix. A previous admission decision for a patient based on certain CCs will not be helpful for the prediction for another patient unless they share common coded CCs. A better method should consider the semantic relatedness between CCs in modeling training and predicting the likelihood of admission. Therefore, we leverage the predictive power of a CC by considering not only its own occurrence, but also the occurrence of related CCs.

Based on this idea, we adapt the standard feature vector representation with enhanced semantics by a smoothing approach. In the CC coding scheme, if two CCs share the same type or system, they are likely to be related. For example, two CCs, “chest pain” and “cardiorespiratory arrest,” share the same type “symptom” and the same system “cardiovascular.” Obviously, these two complaints are more similar to each other than two other complaints that do not belong to the same type or system. To capture such semantic similarity between CCs, we adjust the value of a CC feature based on those of its related CCs. Specifically, a CC affects the value of a related CC of the same type by an impact factor of α ($0 < \alpha < 1$), and a CC affects the value of a related CC of the same system by an impact factor of β ($0 < \beta < 1$). The five-level acuity rating of a coded CC is also considered when we adjust the value of a CC feature. Finally, since each patient record may contain multiple complaints, an individual’s overall CC value should be defined as a weighted sum of effects of all chief complaints assigned to the patient. If a CC C_k (k is the index of C_k) does occur in a patient record ($C_k = 1$), no adjustment is performed. Only if a CC C_k does not occur in a patient record ($C_k = 0$), the value of this feature C_k' was adjusted by the following function (1):

$$C_k' = \frac{\sum_{i=1}^N \frac{\alpha T(i,k)C_i}{r_i} + \sum_{j=1}^N \frac{\beta S(j,k)C_j}{r_j}}{\sum_{i=1}^N T(i,k) + \sum_{j=1}^N S(j,k)} \tag{1}$$

where N is the total number of coded CCs, $T(i, k)$ is a match function to test if C_i and C_k belong to the same type,

$S(j, k)$ is a match function to test if C_j and C_k belongs to the same system, and r_i (r_j) is the acuity rating of C_i (C_j).

Let us consider a patient with a chief complaint “chest pain.” Given the basic feature vector representation, each CC can only take a Boolean value. Thus, except for the feature “chest pain” equal to 1, all remaining CCs will be 0, no matter they share a common system or type with “chest pain” or not. By contrast, after applying the smoothing function, each CC that share a common system or type with “chest pain” will take a value greater than 0. For instance, “cardiorespiratory arrest” has the same type “symptom” and the same system “cardiovascular” as “chest pain.” According to formula (1), its value will not be 0 but a value between 0 and 1. We call the new feature vector with adjusted values a semantic-enhanced feature vector. As such, we embed the semantic meanings of CCs into the feature vector representation, which can potentially improve the performance of prediction model.

3.2.2 Semantic kernel method

In addition to the semantic-enhanced feature method, we also develop an alternative, a kernel method for data representation and model building. Unlike feature methods, kernel methods can capture complex structural information in data points without explicitly elaborating all of the features in a vector. Furthermore, with kernel methods, data in different formats can be easily combined together for model learning. In this study, in order to integrate ED patient data (e.g., demographic, clinical, and CCs) in different formats for admission prediction, we define a new semantic kernel function $K(P_1, P_2)$ that computes the similarity between two patients, P_1 and P_2 . This kernel function K is a composite kernel combining two sub-kernel functions K_L and K_C :

$$K(P_1, P_2) = \lambda K_L(P_1, P_2) + (1 - \lambda) K_C(P_1, P_2) \tag{2}$$

where sub-kernel K_L deals with features other than CCs, K_C deals with CC features, and $0 < \lambda < 1$ represents the weight of the K_L sub-kernel. K_L and K_C transform the training data, including different types of variables, into the same format of kernel matrices.

Specifically, the sub-kernel K_L handles variables regarding patients’ demographic information and clinical information. Since these variables can be represented in the same format, we choose not to further separate them into different sub-kernel functions for the sake of simplicity. Different existing kernel functions can be used for K_L . For example, we can adopt a simple linear kernel function that sums up the similarity score of each feature for the two patients.

$$K_L(P_1, P_2) = P_1 \times P_2 = \sum_u M_u(P_1, P_2) \tag{3}$$

where $M_u(P_1, P_2)$ is a match function for feature u :

$$M_u(P_1, P_2) = \begin{cases} 1 & \text{if } u_1 = u_2 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

For sub-kernel K_C , we define a series of functions to compute the similarity between complaints and between patients. Since each CC C_i contains sub-features such as name w_i , type t_i , system s_i , and acuity rating r_i , we define the similarity function between two CCs as follows:

$$S(C_i, C_j) = \sum_v S_v(C_i, C_j) \quad (5)$$

where $S_v(C_i, C_j)$ gives the similarity between C_i and C_j for the sub-feature v .

Specifically, each complaint’s name w_i is phrase of multiple terms (e.g., “chest pain”). We treat each name w_i as a “bag of words” and thus use the cosine similarity function to compute the similarity between two complaint names w_i and w_j :

$$S_w(C_i, C_j) = \frac{w_i \cdot w_j}{\|w_i\| \|w_j\|} \quad (6)$$

For instance, the cosine similarity between two complaint names, “chest pain” and “abdominal pain,” will be $\frac{1}{\sqrt{1^2+1^2}\sqrt{1^2+1^2}} = \frac{1}{\sqrt{2}\sqrt{2}} = \frac{1}{2}$.

For the CC’s type (t), system (s), and acuity rating (r), the similarity can be computed by a simple match function:

$$S_v(C_i, C_j) = M_v(C_i, C_j) = \begin{cases} 1 & \text{if } v_i = v_j \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where $v \in \{t, s, r\}$.

Each patient record may contain multiple complaints. The similarity between two sets of complaints, $\{C_{11}, C_{12}, \dots, C_{1n_1}\}$ for a patient P_1 and $\{C_{21}, C_{22}, \dots, C_{2n_2}\}$ for a patient P_2 , can be computed by integrating the similarity scores of each individual complaint pairs:

$$K_C(P_1, P_2) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} S(C_{1i}, C_{2j}) \quad (8)$$

where n_1 and n_2 are the numbers of occurring CCs for patients P_1 and P_2 , respectively.

Finally, the kernel matrix needs to be normalized as follows:

$$K * (P_1, P_2) = \frac{K(P_1, P_2)}{\sqrt{K(P_1, P_1)K(P_2, P_2)}} \quad (9)$$

3.3 Learning and evaluation

Once the data transformation phase was done, original patient records from the training data set were either represented as semantic-enhanced feature vectors or transformed into a semantic kernel matrix of similarity scores.

Machine learning algorithms such as an SVM (Cristianini and Shawe-Taylor 2000) can take either the feature vectors or the kernel matrix as input for model training. The trained classification models were then validated using a separate test set based on standard evaluation metrics, such as accuracy, sensitivity, and specificity, compared to other benchmark methods. A model that has been validated by these methods can then be used for predicting the admission of future patients based on their pre-admission information.

4 Experiments

In this study, we conducted experiments on a real ED data set to examine our proposed approaches for hospital admission prediction.

4.1 Data description

The test bed we used in our experiments was an ED data set from Hahnemann Hospital, an academic hospital and level I trauma center serving an inner city population in Philadelphia, Pennsylvania. The ED at Hahnemann serves over 34,000 patients per year and is an approved residency training site for emergency medicine. Triage may occur at bedside or outside the main ED. This study was approved by the Drexel University Institutional Review Board under the expedited classification. A convenience sample of triage data for 2,794 patient visits to the ED of Hahnemann in January 2008 was obtained by extracting data from handwritten charts.

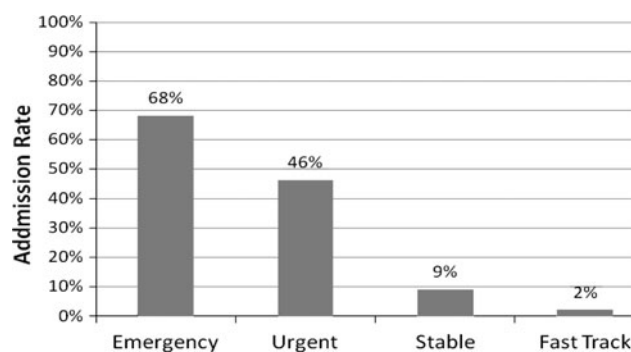
In this data set, information about each patient includes: age, gender, date and time of arrival, an acuity rating assigned by triage nurse (4 levels: emergent, urgent, stable, and fast track), chief complaints (a transcription and possible summarization by the nurse based on a patient’s presenting statements), and, most importantly, decision of admission made by the hospital (whether the patient was admitted for inpatient care). Based on suggestions by a doctor affiliated with the hospital, we preprocessed the raw data to construct variables to address suspected differences in ED and hospital resources as well as patient arrival patterns. Elderly patients are more likely to have a greater number of pre-existing medical problems. For example, age is an independent risk factor for the development of cardiovascular disease and age is associated with greater physiological derangement. It often takes more work and time to help get an elderly patient back to a basic level of health and thus an admission is more likely to be needed. Hence, age was dichotomized to elderly and non-elderly dividing at 60 years of age. Furthermore, due to limited hospital resources, patient arrival patterns in date and time

Table 2 A summary of variables in our ED data set

Variables	Values	Counts
Age (≥ 60)	False	2,321
	True	473
Gender	Male	1,450
	Female	1,344
Weekend	False	2,157
	True	637
Night_shift	False	2,119
	True	457
Acuity	Emergent	85
	Urgent	1,377
	Stable	1,005
	Fast track	325
CCs (top 5)	Chest pain	271
	Abdominal pain	270
	Injury–trauma NOS	217
	Vomiting	215
	Cough	208
Admission	False	2,013
	True	781

may also affect the capacity of hospitals in accommodating new patients who need inpatient care. Date information was used to generate a Boolean variable “weekend.” Time information was used to generate a Boolean variable “night_shift” (i.e., a night shift is between 2300 and 0659 hours). Because chief complaints are available as free text and may vary due to misspelling, abbreviations, and synonyms, the chief complaint information from paper charts was preprocessed by CCC-EDS and mapped to a coded CC (Thompson et al. 2006). Admission information was obtained from charts and was regardless of destination or how long the patient may have stayed in the ED. Among these 2,794 patients, 781 (28 %) were admitted to the hospital. To retain the prior probabilities of the two classes, we did not do any balancing on the data in model training. Table 2 summarizes all variables/features in our data set.

We conducted Chi-square tests to assess the relevance to the target variable “admitted” for each feature, except for the CCs. Results showed that all the features, i.e., gender, age, shift in day, night shift, and acuity, were statistically significant at level $\alpha = 0.05$. In particular, it is not a surprise that acuity rating assigned by nurses seems to be a good predictor for hospital admission. Figure 2 shows the admission rates for the four acuity ratings. In our data set, 68 % of patients with acuity rating as “emergent” were eventually admitted at hospital, whereas only 2 % of patients with acuity rating as “fast track” were admitted. However, it is unreliable to use acuity rating as the only criterion to predict admission. Specifically, if we predict all

**Fig. 2** Admission rate for patients with different acuity ratings

“emergent” patients to be admitted and all others not, the overall accuracy is only 73.16 %. Therefore, we will need to include other features to build a more accurate predictive model for hospital admission.

4.2 Experimental design

In our experiments, we used the decisions of admission made by the hospital as the gold standard for labeling our data points for training and testing. It is worth emphasizing that our research is not aimed at rectifying hospitals’ decisions on patient’s admission. Instead, our main goal is to help hospitals predict patient admissions more timely, so that necessary operational preparations can start earlier and patients’ waiting times can be reduced. Therefore, we assume all decisions for admission made by the hospital were correct and train our prediction models based on them.

In our experiments, we evaluated the performance of our proposed models by comparing them with several classification techniques. Our selected benchmark methods included logistic regression (LR), Naïve Bayes classifier (NB), decision trees (DT), and SVMs. Logistic regression is a general linear model that can predict the likelihood of the occurrence of an event by fitting data to a logistic curve. This statistical technique is widely used in the medical and social sciences for prediction. Naïve Bayes classifier is a simple probabilistic classifier based on Bayes theorem (Han and Kamber 2006). A decision tree is a decision support tool that builds a tree-like graph model of decisions (Han and Kamber 2006). An SVM classifier constructs a separated hyperplane in an n -dimensional space to maximize the margin between the two data groups (Cristianini and Shawe-Taylor 2000). These classifiers were selected because they are commonly used in many applications and show good performances. To demonstrate the predictive power of our semantic-enhanced models, for the benchmark methods, we represented each patient visit as feature vector in which a CC was simply a dichotomous variable indicating its occurrence, with no CC semantics

captured. We did not use one single nominal variable to represent the primary chief complaint as categorical variables because we do not want to lose the information about patients with more than one CC. For both of our proposed approaches, semantic-enhanced feature vector and semantic kernel, we used an SVM to train a classification model. We refer to these two models as SF-SVM and SK-SVM, respectively. For the semantic-enhanced feature method, we assumed that type and system were equally important for representing semantics and tuned the two parameters, α and β , from 0.1 to 0.9 for the best prediction accuracy. Our tuning results showed that SF-SVM achieved its best performance when $\alpha = \beta = 0.5$. For the semantic kernel method, we tuned the parameter λ from 0.1 to 0.9 and found that SK-SVM achieved its best accuracy when $\lambda = 0.5$. In our analysis, we only compare the experimental results under the best parameter settings.

In order to evaluate our proposed admission prediction models, we chose to use two data-mining packages, WEKA (<http://www.cs.waikato.ac.nz/ml/weka/>) and LibSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>) (Hsu et al. 2010). We did not use other commercial software because these two are available for free and, more importantly, they both are used in many data-mining studies for evaluation. WEKA is a Java-based machine learning package that supports a number of standard data-mining tasks. Instead of using the LibSVM module available in the latest WEKA package, we used LibSVM which was chosen for SVM classifiers because it supports customized kernels and parameter selection. For the two classifiers, SVM and SF-SVM, a radial basis function (RBF) kernel was chosen for model learning because this kernel function can handle potentially non-linear relationships between features and target values. In our experiments, we used tenfold cross-validation to evaluate the performance of each classification model because the cross-validation procedure can prevent the problem of overfitting (Hanley and McNeil 1982). Standard evaluation metrics, such as accuracy, sensitivity, specificity, and receiver operating characteristic (ROC) curves, were used to assess the correctness of classification.

4.3 Experimental results

The performances of different admission prediction models are summarized in Table 3. The best performances for the three evaluation metrics are highlighted in bold font. Our two proposed approaches, SF-SVM and SK-SVM, outperformed other benchmark methods in terms of accuracy and sensitivity. Specifically, the semantic-enhanced feature-based approach (SF-SVM) achieved the highest accuracy and sensitivity, 81.21 and 68.03 %, but the lowest specificity, 81.40 %. The semantic kernel approach (SK-

Table 3 Performance of admission prediction models

Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
LR (logistic regression)	77.34	62.25	82.24	0.8337
NB (Naïve Bayes)	77.38	62.34	83.03	0.8088
DT (decision trees)	76.21	60.01	81.93	0.8040
SVM (support vector machine)	78.21	63.12	82.86	0.8234
SF-SVM	81.21	68.03	81.40	0.8388
SK-SVM	79.32	63.21	85.93	0.8808

SVM) achieved the second highest accuracy and sensitivity, 79.32 and 63.21 %, and also the highest specificity, 85.93 %. We have conducted paired *t* tests to compare the performance of different methods. Our tests showed that the accuracy and sensitivity of SF-SVM and the specificity of SK-SVM are statistically higher than those by the baseline methods (*p* value <0.05).

We also analyzed the performance of our admission decision-making methods by comparing ROC curves. In Fig. 3, each ROC curve plots the true positive rate (i.e., sensitivity) versus the false positive rate (i.e., 1 – specificity) at different decision thresholds. The closer the ROC curve is to the upper left corner, the higher the overall accuracy. The area under the ROC curve (AUC) of each classification is another measure to evaluate the performance of the model. As shown in Fig. 3, the ROC curves of SK-SVM and SF-SVM show better overall prediction performance compared to benchmark methods. Table 3 summarizes the area under curve (AUC) values for all prediction models. Of all models, SK-SVM and SF-SVM achieved the highest AUC, 0.8808 and 0.8388, respectively. Statistical tests showed that SK-SVM significantly outperformed other benchmark methods, except for SF-SVM (*p* value <0.05) (Hanley and McNeil 1982).

Our experimental results show that by incorporating the semantic information of CCs, we improved the performance of hospital admission prediction compared to benchmark methods. The semantic-enhanced feature method SF-SVM tended to categorize more patients as requiring admission and therefore gave higher sensitivity (more true positives), but only at the cost of a lower specificity (more false positives). In contrast, although the semantic kernel method SK-SVM did not perform as well as SF-SVM in terms of accuracy and sensitivity, it was able to improve the prediction performance without increasing false positives and therefore is less likely to cause the waste of hospital resources. The primary success of any admission prediction model involves its ability to maximize accuracy. The application of prediction model to the

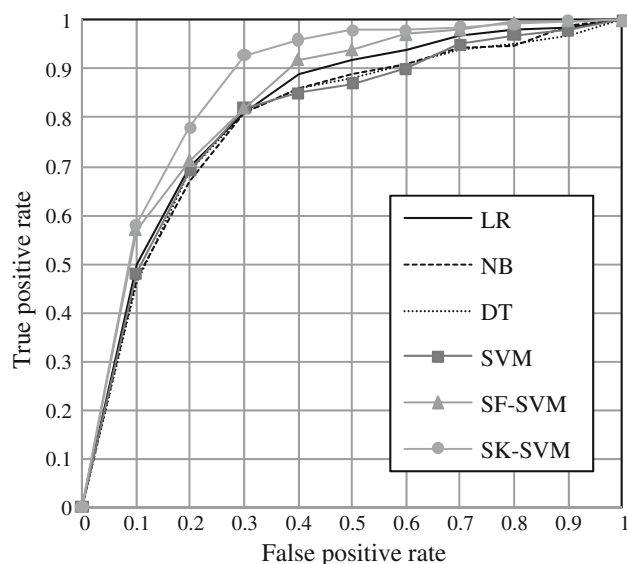


Fig. 3 ROC curves of prediction models

admission process should also consider its effects on costs of and quality of care, which we plan to address in our future research.

For our baseline classification models such as logistic regression and decision trees, the predictive powers of variables can be easily interpreted. In these models, not surprisingly, the acuity rating assigned by the nurse at EDs always showed to be a good predictor variable for admission prediction. However, without capturing the semantics of CCs, the sparse data with hundreds of CC features did not seem to contribute a lot to the prediction. Furthermore, we examined the prediction outcomes and found several examples in which semantic similarity between CC terms rectified the false predictions of models based on basic feature vector representation. For example, a female patient who arrived at night shift was assigned a chief complaint “sores” and acuity rating = “fast track.” The benchmark methods incorrectly predicted her to be admitted to the hospital. By contrast, both of our proposed approaches, SF-SVM and SK-SVM, predicted correctly by recognizing that this patient’s CC, “sores”, shares the same system, “skin, hair, nails, and breast,” and the same type, “symptom,” with another CC, “rash,” which was assigned to another similar patient in the training set who was not admitted to the hospital. For another, a patient with a CC “chest pain” and acuity rating = “emergent,” the benchmark methods categorized her as “not admitted.” By contrast, SF-SVM and SK-SVM captured the semantic similarity between “chest pain” and another CC, “syncope,” which shares the same system, “cardiovascular,” and the same type, “symptom.” Furthermore, because another emergent patient with “syncope” in the training set was admitted to the hospital, our two models rectified the false negative

prediction and correctly categorized the emergent patient with “chest pain” to be admitted as well. These examples showed that, by capturing semantic similarity between complaints, our proposed models can predict admissions more accurately.

5 Conclusions and future directions

Facing the conflicts between the limited hospital resources and the increasing demands of patients, hospitals need to plan and manage care in more effective and efficient ways. To solve the problem of ED crowding, in this study, we proposed two novel classification models that can capture semantic information about chief complaints for hospital admission prediction. We evaluated our prediction models on a 1-month data set collected from an emergency department by comparing them with several benchmark methods. Our experiments showed an encouraging improvement of prediction performance for the proposed approaches in terms of accuracy, sensitivity, and specificity. Our prediction models can potentially be used as decision support tools at hospitals to improve the ED throughput rate and enhance the quality of patient care.

Our study has the following limitations: triage data does not include nurse or physician gestalt about the need for admission. Additionally, there is no information available about the likelihood that a patient might need admission because of an inability to care for oneself at home (social factors). The CCA has been shown to describe case mix but not been validated as a tool that can provide operational power. It may cause important details describing the patient’s condition to be lost. This was a small sample of ED data and it will be important to test these techniques on larger more comprehensive data sets, so that more variables can be used. It was not possible to identify periodic patterns (such as time of month or seasonal factors). The collecting of more factors does come at a cost of increased labor to extract these elements until such information is automatically and regularly collected with electronic data systems.

In our future work, we plan to refine our model design by incorporating other knowledge recourses such as the ICD-9-CM coding scheme and validate our models on larger data sets. The value of our admission model could be enhanced with the inclusion of cost and revenue functions. We will also consider both the costs of false positives and negatives and the opportunity costs of being unable to care for additional patients in the ED if admitted patients cannot be moved quickly. The costs associated with altered quality of care due to ED crowding must also be considered (Bernstein et al. 2009; Pines et al. 2006). Furthermore, we plan to develop a system that can operationalize our

proposed models to support better hospital operations management and patient care.

References

- Accounting G (2003) HOSPITAL EMERGENCY Crowded Conditions Vary among Hospitals and Communities. *Time* (March)
- Achour SL, Dojat M, Rieux C, Bierling P, Lepage E (2001) A UMLS-based knowledge acquisition tool for rule-based clinical decision support system development. *J Am Med Inform Assoc* 8(4): 351–360
- Almansoori W, Gao S, Jarada TN, Elsheikh AM, Murshed AN, Jida J, Alhaji R et al (2012) Link prediction and classification in social networks and its application in healthcare and systems biology. *Netw Model Anal Health Inform Bioinform*. doi:10.1007/s13721-012-0005-7
- Bernstein SL, Aronsky D, Duseja R, Epstein S, Handel D, Hwang U, McCarthy M et al (2009) The effect of emergency department crowding on clinically oriented outcomes. *Acad Emerg Med* 16(1):1–10
- Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 32(Database issue):D267–D270
- Chalfin DB, Trzeciak S, Likourezos A, Baumann BM, Dellinger RP (2007) Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Crit Care Med* 35(6):1477–1483. doi:10.1097/01.CCM.0000266585.74905.5A
- Chan T, Arendts G, Stevens M (2008) Variables that predict admission to hospital from an emergency department observation unit. *Emerg Med Australas* 20(3):216–220. doi:10.1111/j.1742-6723.2007.01043.x
- Chapman WW, Dowling JN, Wagner MM (2005) Classification of emergency department chief complaints into 7 syndromes: a retrospective analysis of 527,228 patients. *Ann Emerg Med* 46(5):445–455
- Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods, vol 3. Cambridge University Press, Cambridge, pp 1–17. doi:10.2277/0521780195
- Derlet RW, Richards JR (2000) Overcrowding in the nation's emergency departments: complex causes and disturbing effects. *Ann Emerg Med* 35(1):63–68
- Derlet R, Richards J, Kravitz R (2001) Frequent overcrowding in U.S. emergency departments. *Acad Emerg Med* 8(2):151–155
- Dickinson G (1989) Emergency department overcrowding. *Can Med Assoc J* 140(3):270–271
- Emergency Severity Index (ESI) (2012) Retrieved from <http://www.ahrq.gov/research/esi/>
- Gabrilovich E, Markovitch S (2007) Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: Proceedings of the 20th international joint conference on artificial intelligence, vol 7. Morgan Kaufmann Publishers Inc, San Francisco, pp 1606–1611
- Haas SW, Travers D, Tintinalli JE, Pollock D, Waller A, Barthell E, Burt C et al (2008) Toward vocabulary control for chief complaint. *Acad Emerg Med* 15(5):476–482
- Han J, Kamber M (2006) Data mining: concepts and techniques. Morgan Kaufmann, San Francisco, p 770
- Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1):29–36
- Herbert I, Hawking M (2005) Bringing SNOMED-CT into use within primary care. *Inform Prim Care* 13(1):61–64
- Hsu C-W, Chang C-C, Lin C-J (2010) A practical guide to support vector classification. *Bioinformatics* 1(1):1–16
- Hulse J, Khoshgoftaar TM, Napolitano A, Wald R (2012) Threshold-based feature selection techniques for high-dimensional bioinformatics data. *Netw Model Anal Health Inform Bioinform*. doi:10.1007/s13721-012-0006-6
- ICD-9 (2012) ICD-9 diagnosis code reference chart. <http://www.cdc.gov/nchs/icd/icd9.htm>
- Lanckriet GRG, De Bie T, Cristianini N, Jordan MI, Noble WS (2004) A statistical framework for genomic data fusion. *Bioinformatics* 20(16):2626–2635
- Leacock C, Chodorow M (1998) Combining local context and WordNet similarity for word sense identification. In: Fellbaum C (ed) *WordNet an electronic lexical database*, vol 49. MIT Press, Cambridge, pp 265–283
- Leegon J, Jones I, Lanaghan K, Aronsky D (2005) Predicting hospital admission for Emergency Department patients using a Bayesian network. In: AMIA annual symposium proceedings. American Medical Informatics Association, Washington, DC, p 1022
- Leegon J, Jones I, Lanaghan K, Aronsky D (2006) Predicting Hospital Admission in a Pediatric Emergency Department using an Artificial Neural Network. In: AMIA annual symposium proceedings. American Medical Informatics Association, Washington, DC, p 1004
- Lewis RJ, Street WC (2000) An introduction to classification and regression tree (CART) analysis. In: 2000 Annual Meeting of the Society for Academic Emergency Medicine, vol 310, p 14
- Li J, Su H, Chen H, Futscher BW (2007) Optimal search-based gene subset selection for gene array cancer classification. *IEEE Trans Inf Technol Biomed* 11(4):398–405
- Li J, Zhang Z, Li X, Chen H (2008) Kernel-based learning for biomedical. *J Am Soc Inform Sci* 59(5):756–769. doi:10.1002/asi
- Lord PW, Stevens RD, Brass A, Goble CA (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19(10):1275–1283
- Lu H-M, Zeng D, Trujillo L, Komatsu K, Chen H (2008) Ontology-enhanced automatic chief complaint classification for syndromic surveillance. *J Biomed Inform* 41(2):340–356
- Lyon D, Lancaster GA, Taylor S, Dowrick C, Chellaswamy H (2007) Predicting the likelihood of emergency admission to hospital of older people: development and validation of the Emergency Admission Risk Likelihood Index (EARLI). *Fam Pract* 24(2):158–167. doi:10.1093/fampra/cml069
- McCarthy D, Koeling R, Weeds J, Carroll J (2004) Finding Predominant word senses in untagged text. In: Proceedings of the 42nd annual meeting on association for computational linguistics ACL 04, 40, 279-es. Association for Computational Linguistics
- Miller G (1990) WordNet: an on-line lexical database. *Int J Lexicogr* 3(4):235–244
- Miró O, Antonio MT, Jiménez S, De Dios A, Sánchez M, Borrás A, Millá J (1999) Decreased health care quality associated with emergency department overcrowding. *Eur J Emerg Med* 6(2):105–107
- Patwardhan S, Banerjee S, Pedersen T (2003) Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the fourth international conference on intelligent text processing and computational, vol 2276. Springer, Berlin, pp 241–257. doi:10.1007/3-540-36456-0
- Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG (2007) Measures of semantic similarity and relatedness in the biomedical domain. *J Biomed Inform* 40(3):288–299
- Pines JM, Hollander JE, Localio AR, Metlay JP (2006) The association between emergency department crowding and

- hospital performance on antibiotic timing for pneumonia and percutaneous intervention for myocardial infarction. *Acad Emerg Med* 13(8):873–878
- Pitts SR, Niska RW, Xu J, Burt CW (2008) National Hospital Ambulatory Medical Care Survey: 2006 emergency department summary. *Natl Health Stat Report* 7:1–38
- Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. *IEEE Trans Syst Man Cybern* 19(1):17–30. doi:[10.1109/21.24528](https://doi.org/10.1109/21.24528)
- Richardson DB (2006) Increase in patient mortality at 10 days associated with emergency department overcrowding. *Med J Aust* 184(5):213–216
- Sadeghi S, Barzi A, Sadeghi N, King B (2006) A Bayesian model for triage decision support. *Int J Med Inform* 75(5):403–411
- Schneider D, Appleton L, McLemore T (1979) A reason for visit classification for ambulatory care. Series 2, data evaluation and methods research. *Vital Health Stat* 2(78):i–vi, 1–63
- Shapiro AR (2004) Taming variability in free text: application to health surveillance. *MMWR Morb Mortal Wkly Rep* 53(Suppl): 95–100
- Sprivulis PC, Da Silva J-A, Jacobs IG, Frazer ARL, Jelinek GA (2006) The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. *Med J Aust* 184(5):208–212
- Steele R, Kiss A (2008) EMDOC (Emergency Department overcrowding) Internet-based safety net research. *J Emerg Med* 35(1):101–107
- Steele R, Green SM, Gill M, Coba V, Oh B (2006) Clinical decision rules for secondary trauma triage: predictors of emergency operative management. *Ann Emerg Med* 47(2):135
- Thompson DA, Eitel D, Fernandes CMB, Pines JM, Amsterdam J, Davidson SJ (2006) Coded chief complaints—automated analysis of free-text complaints. *Acad Emerg Med* 13(7):774–782
- Travers DA, Haas SW (2004) Evaluation of emergency medical text processor, a system for cleaning chief complaint text data. *Acad Emerg Med* 11(11):1170–1176
- Ureña-López LA, Buenaga M, Gomez JM (2001) Integrating linguistic resources in TC through WSD. *Comput Humanit* 35(2):215–230
- Wilbur WJ, Yang Y (1996) An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput Biol Med* 26(3):209–222
- Wilper AP, Woolhandler S, Lasser KE, McCormick D, Cutrona SL, Bor DH, Himmelstein DU (2008) Waits to see an emergency department physician: US trends and predictors, 1997–2004. *Health Aff* 27(2):w84–w95
- Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N (2000) Reliability and validity of a new five-level triage instrument. *Acad Emerg Med* 7(3):236–242
- Zhou S, Wang K (2005) Localization site prediction for membrane proteins by integrating rule and SVM classification. *IEEE Trans Knowl Data Eng* 17(12):1694–1705. doi:[10.1109/TKDE.2005.201](https://doi.org/10.1109/TKDE.2005.201)