Journal of Biomedical Informatics 59 (2016) 209–217

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Classification of hospital acquired complications using temporal clinical information from a large electronic health record

Jeremy L. Warner [a,b,*], Peijin Zhang [c], Jenny Liu [d], Gil Alterovitz [d,e,f]

[a] Department of Medicine, Division of Hematology & Oncology, Vanderbilt University, Nashville, TN, USA
[b] Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA
[c] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[d] Department of Electrical Engineering & Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA
[e] Center for Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[f] Children's Hospital Informatics Program at Harvard-MIT Health Sciences & Technology, Boston, MA, USA

ARTICLE INFO

*Article history:*
Received 4 May 2015
Revised 18 November 2015
Accepted 12 December 2015
Available online 17 December 2015

*Keywords:*
Hospital acquired complications
Electronic health record
Classification
Temporal
Ranking

ABSTRACT

Hospital acquired complications (HACs) are serious problems affecting modern day healthcare institutions. It is estimated that HACs result in an approximately 10% increase in total inpatient hospital costs across US hospitals. With US hospital spending totaling nearly $900 billion per annum, the damages caused by HACs are no small matter. Early detection and prevention of HACs could greatly reduce strains on the US healthcare system and improve patient morbidity & mortality rates. Here, we describe a machine-learning model for predicting the occurrence of HACs within five distinct categories using temporal clinical data. Using our approach, we find that at least $10 billion of excessive hospital costs could be saved in the US alone, with the institution of effective preventive measures. In addition, we also identify several keystone features that demonstrate high predictive power for HACs over different time periods following patient admission. The classifiers and features analyzed in this study show high promise of being able to be used for accurate prediction of HACs in clinical settings, and furthermore provide novel insights into the contribution of various clinical factors to the risk of developing HACs as a function of healthcare system exposure.



## 1. Introduction

Hospital acquired complications (HACs) are secondary complications that affect patients following initial hospital admission. Most commonly, these problems are caused by healthcare-associated infections and other issues resulting as side effects from primary treatments such as surgery. Although well known to occur, many HACs are not identified in a timely manner, and often lead to further medical issues causing increased hospitalization time, temporary and permanent morbidities, or even death. It has been estimated that HACs alone cause nearly 99,000 deaths annually in the United States [1] and lead to an approximately 10% increase in total inpatient hospital costs [2]. Early identification of HACs using electronic health records (EHRs) may lead to prevention or proactive treatment, ultimately increasing patient care quality and reducing unnecessary inpatient costs.

The development and deployment of large-scale EHR databases has led to a surge in medical informatics over the past several years [3–6]. In particular, phenotypes across the human "phenome" have been shown to have significant associations with genotype and with lab features [7–9]. Since the manifestation of a HAC is also a patient phenotype, it follows that there may be a way to identify and predict HACs using temporal information that unfolds during a hospitalization. We have previously shown that certain patient phenotypes, many of which are HACs, are associated with prolonged hospitalization [10].

The Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II) database contains extensive structured clinical information on patients admitted to any of the Beth Israel Deaconess Medical Center Intensive Care Units (ICUs) between 2001 and 2007,

*Abbreviations:* HAC, hospital acquired complication; EHR, electronic health record; MIMIC, Multiparameter Intelligent Monitoring in Intensive Care; ICD-9-CM, International Classification of Diseases, 9th Edition, Clinical Modification; ICU, Intensive Care Unit; TPR, true positive rate; TNR, true negative rate; ROC, receiver operating characteristic curve; AUC, area under the ROC curve.

* Corresponding author at: Medicine and Biomedical Informatics at Vanderbilt University, 2220 Pierce Ave, Preston Research Building 777, Nashville, TN 37232, USA. Tel.: +1 (615) 322 5464; fax: +1 (615) 343 7602.
*E-mail address:* jeremy.warner@vanderbilt.edu (J.L. Warner).

http://dx.doi.org/10.1016/j.jbi.2015.12.008
1532-0464/

including details of their hospital stay before and after the ICU component [11]. MIMIC II also contains some narrative information: nursing documentation, radiology reports, and Discharge Summaries. Here, we propose to use MIMIC II to develop predictive models to detect HACs early on in hospitalization, possibly in advance of their occurrence. As a secondary objective, we hypothesized that the relative contribution of predictor variables (e.g. patient age, lab results, and medication orders) would change as a function of the timespan used to generate the predictive model. Analysis of this rich medical cohort can lead to new discoveries for the prediction and possible early prevention of problematic HACs.

## 2. Methods

### 2.1. Data extraction and modeling

To identify HACs in the MIMIC II cohort, we used a candidate list of International Classification of Diseases, 9th Edition, Clinical Modification (ICD-9-CM) codes that were previously determined to be associated with an increased risk of prolonged hospitalization [10]. These and additional ICD-9-CM codes were manually reviewed and all codes that were indicative of HACs were used to identify patients having HACs. A total of 414 ICD-9-CM codes were selected through this process. We further categorized each identified ICD-9-CM code into five subcategories: (1) infectious complications; (2) bleeding or clotting complications; (3) surgical/procedural complications; (4) medical complications; and (5) all other complications (Table A1). In order to decrease the possibility that a candidate code was representative of a pre-existing condition, we compared our candidate list to the Elixhauser AHRQ-Web ICD-9-CM list of comorbidities [12,13]. There were 2 overlaps (287.41: post-transfusion purpura & 276.61: transfusion with circulatory overload), but neither of these codes was present in the MIMIC II cohort. We also manually reviewed the Discharge Summaries from 50 randomly selected admissions from each subcategory to determine whether the complication was the result of a prior healthcare event (e.g. a preceding outpatient procedure, or a previous hospitalization).

All adult MIMIC II hospitalizations having one or more of these identified diagnosis codes assigned at discharge were considered cases, and an equal number of hospital admissions were randomly sampled from the remaining MIMIC II hospitalizations as controls, excluding those of duration outside the 1st to 99th percentiles of the cases. Patients with more than one hospitalization could be assigned to case or control groups, based on the ICD-9-CM codes obtained for a particular hospital stay.

Structured data was extracted from MIMIC II for both cases and controls, beginning at the time of initial contact, which usually occurs in the outpatient setting (e.g. the emergency room). Initial and last contact of a patient was algorithmically defined based on our prior definition [10]. Clinical information was extracted for each admission based on time intervals of 1, 2, 3, 6, 12, 18, 24, 48, 72, and 96 h following initial contact; patients whose last contact times were outside of the specified interval were censored for that interval. Structured information was extracted by recording the maximum, minimum, and median values for continuous lab results, the mode for categorical lab results, any procedure codes (ever/never), and any medication provider order entry (POE) (ever/never) time-stamped during each interval. In order to increase power, we developed a large library of custom medication aggregations e.g. to combine a medication given by more than one route but with the same predicted pharmacologic effect (Appendix B). Patient demographics (age at admission, gender, and ethnicity) were included as well.

To prevent excessive noise, over-fitting, and unreasonable demands on the prediction model, initial feature selection was carried out as follows: lab features missing in more than 50% and POE/procedure features missing in more than 95% of our cohort were removed prior to classifier construction. Following feature selection, multiple imputation was performed using the aregImpute algorithm to impute missing values (R package Hmisc) [14].

### 2.2. Feature ranking

To assess the predictive power of individual features, we computed the information gain ratio of each attribute relative to the class feature [15]. Attributes with high information gain ratios with respect to the classification attribute (presence of HAC) are likely

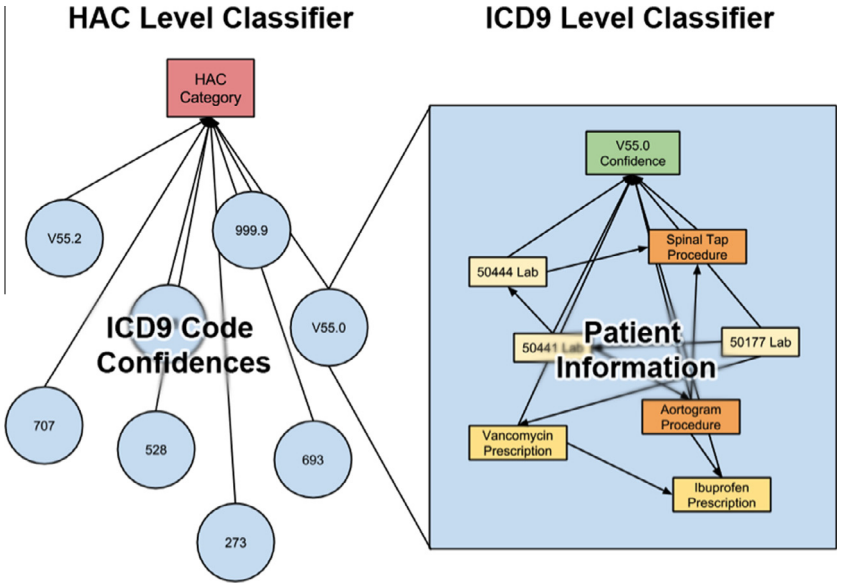


**Fig. 1.** Illustration of two level classifier model used in this study. The bottom level (ICD-9-CM prediction) model is a Bayesian network that uses patient information to produce a probability of observing the ICD-9-CM code. The top level (HAC prediction) model uses confidences given by the first level Bayesian network to perform binary classification on whether or not a given patient falls under a HAC category.