Predicting length of stay and assignment of diagnosis codes during hospital inpatient episodes

José Carlos Ferrão, jose.ferrao@tecnico.ulisboa.pt, Siemens Healthcare
Mónica Duarte Oliveira, monica.oliveira@tecnico.ulisboa.pt, CEG-IST
Filipe Janela, filipe.janela@siemens.com, Siemens Healthcare
Henrique Martins, hmartins@fcsaude.ubi.pt, CI2 - HFF

Electronic health record (EHR) data is becoming ubiquitous in the healthcare domain, with potential to provide valuable insights to clinicians and managers. Data mining methodologies have been largely unexplored to analyze EHR data retrospectively and to inform expected patterns of disease and utilization during the course of patient stay. In this work, we propose a data mining methodology based on feature selection and logistic regression models to predict if an episode’s length-of-stay will be outside the expected interval, as well as the set of diagnosis codes (from the International Classification of Diseases – ICD) assigned to each episode along the course of patient stay. The experiments were carried out using EHR data from the records of 5089 episodes of inpatients admitted in a large (772-bed) hospital in Portugal. The predictive performance of models in terms of precision, recall and F1-score values showed the potential value of using decision support tools during patient stay, since in several experiments the developed models exhibited performance values with sufficient accuracy to provide support in clinical settings.

1 Introduction

In recent years, information systems have undergone extensive development in the healthcare domain (Ford, Menachemi, and Phillips 2006), where EHR systems have played a central role as main platform for recording clinical data. The large volumes of EHR data have induced interest in reusing these data for research and decision support (Hersh 2007) and in increasing the proportion of data captured in structured formats (Fernando et al. 2012) so as to surpass the challenges of using narrative data (Jaspers et al. 2011).

Data mining in medicine has been a hot topic in recent years (Bellazzi and Zupan 2008; Patel et al. 2009), making use of techniques to extract knowledge from data and support clinicians and managers on their decisions (Iavindrasana et al. 2009). The applications include length-of-stay (LOS) (Rowan et al. 2007) and clinical coding (Stanfill et al. 2010) prediction, amongst others, and typically encompass elements such as retrieval and preparation of historical data, defining a set of variables (features), applying feature selection techniques (when suitable)
and building models to extract patterns from data. However, most research adopts a retrospective batch approach whereby models are developed using all data available in the dataset, without evaluating the potential of developing decision support tools during the course of patient care, as more information becomes available. In the case of inpatient episodes, decision support on clinical and managerial elements such as diagnoses and LOS, respectively, during the course of the episode would be valuable to inform timely decisions.

In this scope, we sought to analyze the extent to which decision support tools may be developed to inform expected LOS (which can be regarded as a proxy of resource utilization) and episode coding (which may inform on disease patterns) during the course of the episode. For this purpose, we propose a data mining methodology to build models for predicting LOS and episode coding in different instants of patient stay. In effect, previous LOS prediction studies have analyzed LOS in different moments of patient stay, yet, up to our knowledge, no study has used a systematic approach of defining features and building models consistently across different moments (and comparing performance amongst them). On the other hand, as far as we could appraise, episode coding has only been addressed within a retrospective frame, without addressing coding support during the course of patient stay. The rest of this article is structured as follows: section 2 describes the proposed methodology, section 3 presents key results to illustrate an application with a real world dataset for LOS and episode coding prediction, and lastly section 4 presents our main concluding remarks.

2 Methods

2.1 EHR data properties and preparation

In this work, the dataset from which we set off consists of EHR database entries produced during clinical practice using the EHR system Soarian® (Haux et al. 2003). These entries contain patient information in structured formats, regarding demographic data, diagnoses, personal history, allergies, prescriptions and medication, as well as structured forms (assessments) composed of labeled fields through which health professionals record information using controlled formats (such as buttons, pick lists and dropdowns). The use of such system eliminates the need to extract computer-readable information from narratives (in which information is “locked” (Hripcsak et al. 1995)) and, thereby, the challenges associated with it.

The development of prediction models based upon structured data requires data to be represented in a data matrix format, i.e., with dataset instances represented in terms of values of the feature space (i.e., a matrix in which lines represent instances and columns represent features) (Bishop 2006). Since EHR data is natively represented as relational database entries (and not in a data matrix format), it was necessary to define features from data. In practice, defining features from data consists in defining the variables based on which clinical data can be represented and whose values (for each instance) are used to build prediction models. This feature definition process requires identifying the clinical concepts contained in the dataset, defining a feature for each concept and assessing its value for each instance (episode) in the dataset, thereby building and populating the data matrix.
matrix representation. However, for large datasets (with high number of features and instances), the process of building and populating a data matrix becomes laborious. In order to mitigate this issue, we implemented a framework with a set of routines that build and populate a data matrix automatically from a dataset (further details available in (Ferrão et al. 2013)). Therefore, after using this framework to automatically digest the dataset and build a data matrix from the original EHR database entries, it was then possible to work on the development of prediction models using the clinical dataset represented in a data matrix format.

2.2 Feature selection and model development

In this study, the decision support tools to predict LOS and episode coding were based on prediction models developed from historical data (i.e., clinical records from past episodes) aiming to predict outcomes for new episodes. To this end, we adopted a supervised learning approach (Bishop 2006), which consists in using data from past instances (episodes) and the corresponding known outcomes (in this case, LOS and assigned clinical codes) to fit models that are subsequently used to predict those outcomes for new instances. There are two main types of models within supervised learning models, depending on the type of outcome to be predicted: if the outcome is discrete or nominal, classification models are employed, as opposed to regression models used for continuous outcomes. In our approach, we modeled LOS and episode coding prediction as binary outcomes. Firstly, we modeled LOS as a binary variable according to whether or not the duration of patient stay is within the boundaries of the corresponding diagnosis-related group (DRG) (and thus within the fixed-rate payment scheme). Secondly, for episode coding, we defined a binary variable for each ICD code in the dataset: since multiple codes can be assigned to each episode, we built a model predicting the assignment of each ICD code to a given episode. We hereby describe the processes of data preparation, feature selection and model development.

The data matrix resulting from structured EHR data tends to contain a large number of features, which hinders model performance and, as such, feature selection methods come to play (Guyon and Elisseeff 2003). We chose to implement a filter method for its scalability to large datasets and independence of prediction models (Saey, Inza, and Larrañaga 2007). We performed several tests with different filter methods by evaluating the predictive power of multiple filter methods, namely fast correlation-based filter (Yu and Liu 2004), information gain and chi-square (Yiming Yang 2014), Relief (Kira and Rendell 1992), symmetrical uncertainty (Press et al. 1992), correlation-based feature selection (Hall 1999) and minimal-redundancy maximal-relevance (mRMR) (Peng, Long, and Ding 2005). In these preliminary tests, mRMR revealed higher performance, and as such, we present results obtained with this method in section 3 (also due to space constraints). In practical terms, we used mRMR feature selection to select a subset of features of the original feature set according to the mRMR criterion, which defines a score based on mutual information and aims to minimize redundancy while maximizing relevance of the feature subset. The mRMR method outputs features sorted by decreasing order of mRMR score. From this ordered set, we selected a subset of the top 50 features. This 50-feature subset was then used to develop logistic regression models.
The prediction models used in this study – logistic regression models – have simple formulations, reasonable scalability and interpretability of results, and are tailored for binary outputs (Hosmer and Lemeshow 2000), being used in this work to predict LOS and ICD code assignment. Logistic regression models were developed in a forward selection, stepwise approach, adding one feature at a time by decreasing order of mRMR score. For each logistic regression model developed, we used classification thresholds ranging from 0 to 1 in steps of 0.005 (in order to compensate for class imbalances). Model performance was tested with 5-fold cross-validation, whereby the dataset is randomly split into 5 subsets, using 4 of these subsets as training set and the remaining one as test set. As evaluation metrics, we analyzed precision and recall, which are based on the proportion of false positives and false negatives, respectively, as well as the F1-score, which is the harmonic mean between the precision and recall.

Our experiment approach consisted in (1) extracting structured EHR data, (2) creating 8 separate datasets for each of the moments after patient admission (using the time stamps associated with each database entry), (3) building a data matrix for each separate dataset, (4) performing feature selection with the mRMR method and (5) developing prediction models and evaluating predictive power for each of the 8 datasets. Predictive power was evaluated at 8 different moments of patient stay: 1, 4, 8, 12, 18, 24, 36 and 48 hours after patient admission. For this purpose, we firstly determined the date/time boundaries and used the timestamps in EHR database records to filter data, thereby creating a dataset for each instant.

3 Results

Our real-world dataset contained 5089 inpatient episodes from medical wards of a large hospital in Portugal. EHR data from these episodes was extracted and prepared, yielding 4820 features with non-missing values. The distribution of positive and negative examples was quite imbalanced: 15.72% of positive examples in the LOS problem, one ICD code with 40% positive examples, 5 other codes with more than 15% of positive instances, while the remaining had less than 10% positive examples.

3.1 LOS prediction

In the first experiment, we developed models to predict whether an episode is within the boundaries of the DRG class it is assigned, in different instants of each episode. The results are depicted in Fig. 3.1. As expected, one may observe that model predictive power tends to increase along the course of the episode. This tendency is evident for precision, i.e., the number of false positives decreases as more information becomes available in the EHR. Recall seems to decrease in initial stages of each episode, recovering along the episode. It is also interesting to note that precision was tendentially higher than recall, which in practice yields that these models are more suitable to correctly spot LOS-outlier episodes. The overall performance (in terms of F1-score) has a steeper increasing trend in the first instants, reaching acceptable values (higher than 50%) in early stages of patient stay.
3.2 ICD codes prediction

We also developed logistic regression models to predict the assignment of ICD-9-CM diagnosis codes. In order to keep the analysis manageable, we focused on the 50 more frequent codes. We analyzed performance averaged across all 50 codes and also analyzed performance for selected codes that showed different tendencies (Fig. 3.2).
The trends shown in Fig. 3.2(a) exhibit similarities with the ones obtained with models for LOS prediction, namely in what concerns the increase in average predictive power in initial hours of patient stay. Comparing precision and recall, these measures now exhibit different behaviors, with recall being tendentially higher than precision, rendering models more prone to suggesting incorrect codes while decreasing the rate of overlooked codes. It is also interesting to observe the non-monotonic behavior of recall in initial stages of patient stay, suggesting that the patterns of EHR data produced in such periods may induce biases in models.

In order to further investigate the pool of analyzed ICD codes, we selected codes with different behaviors, which are depicted in Fig. 3.2(b). In this chart, it is possible to identify two different patterns, corresponding to codes exhibiting, or not, a steep increase in performance during the course of the episode. ICD codes referring to pneumonia, cerebral artery occlusion and anemia had a marked tendency to increase predictive power on early stages, while the other selected codes approximately maintained model results. In light of such results, we hypothesize that the potential to develop decision support tools and provide valuable insights during patient stay is highly dependent on the type of outcome being predicted, and especially on the characteristics of EHR data (e.g. data quality, feature subsets) underlying model development.

4 Conclusions and future work

In this study, we proposed a methodology to analyze the extent to which valuable decision support may be provided during the course of the episode, using a real-world dataset of inpatient episodes. Specifically, after data preparation and feature selection, we have built logistic regression models and analyzed model performance in predicting LOS-outliers and ICD code assignment in selected instants of patient stay. Our main conclusions refer to the fact that, as expectable, model performance tends to increase during the course of the episode, especially in the first hours after patient admission. Secondly, this behavior was observed in different magnitudes in predicting LOS and code assignment, which denotes the influence of modeling context and of dataset characteristics in model behavior. It was also possible to observe non-monotonic behaviors (Fig. 3.1 and 3.2(a)), which point to the possibility that statistical artifacts and data quality issues may be exerting influence on model results. Lastly, we could observe that model performance does not always exhibit a steadily increasing trend, either by starting off at higher values (e.g. code 401.9) or maintaining lower values in spite of the increasing availability of EHR data.

From these preliminary results, we wrap up by stating that the research path of developing decision support tools to provide on-the-fly insights in healthcare settings may be promising, as demonstrated by the existence of high-performing models on very early stages of patient stay, with strong correlation with the context of application and the availability of data.

In terms of future work, it should be relevant to firstly carry out a thorough analysis of data quality and data patterns produced in different stages of patient stay, in order to explore the value of information at different
stages of each episode. It may also be relevant to analyze feature subsets in further detail using domain knowledge from clinical experts as a means to improve the set of features that are used to build prediction models. Furthermore, it should also be relevant to test the behavior of different prediction models during the course of the episode. In effect, we have tested other models in the scope of coding support in our previous works (Ferrão et al. 2012; Ferrão et al. 2013), but only testing model behavior using all available EHR data. Lastly, additional methods to tackle issues of inter-label relationships and class imbalance may also be valuable in this context.

References


