

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:<https://orca.cardiff.ac.uk/id/eprint/138309/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Ferrão, José Carlos, Oliveira, Mónica Duarte, Gartner, Daniel , Janela, Filipe and Martins, Henrique 2021. Leveraging electronic health record data to inform hospital resource management: A systematic data mining approach. *Health Care Management Science* 24 , pp. 716-741. 10.1007/s10729-021-09554-4

Publishers page: <http://dx.doi.org/10.1007/s10729-021-09554-4>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



Leveraging electronic health record data to inform hospital resource management

A systematic data mining approach

José Carlos FERRÃO · Mónica Duarte OLIVEIRA · Daniel GARTNER · Filipe JANELA · Henrique M. G. MARTINS

Initial submission: May 26, 2020; 1st revision: December 8, 2020; 2nd revision: February 6, 2021

Abstract Early identification of resource needs is instrumental in promoting efficient hospital resource management. Hospital information systems, and electronic health records (EHR) in particular, collect valuable demographic and clinical patient data from the moment patients are admitted, which can help predict expected resource needs in early stages of patient episodes. To this end, this article proposes a data mining methodology to systematically obtain predictions for relevant managerial variables by leveraging structured EHR data. Specifically, these managerial variables are: i) Diagnosis categories, ii) procedure codes, iii) diagnosis-related groups (DRGs), iv) outlier episodes and v) length of stay (LOS). The proposed methodology approaches the problem in four stages: Feature set construction, feature selection, prediction

model development, and model performance evaluation. We tested this approach with an EHR dataset of 5,089 inpatient episodes and compared different classification and regression models (for categorical and continuous variables, respectively), performed temporal analysis of model performance, analyzed the impact of training set homogeneity on performance and assessed the contribution of different EHR data elements for model predictive power. Overall, our results indicate that inpatient EHR data can effectively be leveraged to inform resource management on multiple perspectives. Logistic regression (combined with minimal redundancy maximum relevance feature selection) and bagged decision trees yielded best results for predicting categorical and numerical managerial variables, respectively. Furthermore, our temporal analysis indicated that, while DRG classes are more difficult to predict, several diagnosis categories, procedure codes and LOS amongst shorter-stay patients can be predicted with higher confidence in early stages of patient stay. Lastly, value of information analysis indicated that diagnoses, medication and structured assessment forms were the most valuable EHR data elements in predicting managerial variables of interest through a data mining approach.

José Carlos FERRÃO and Filipe JANELA
SIEMENS Healthineers
Rua Irmãos Siemens 1
2720-093 Amadora, Portugal

Mónica Duarte OLIVEIRA
CEGIST, Centre for Management Studies of Instituto Superior Técnico, Universidade de Lisboa
University of Lisbon
Av. Rovisco Pais 1
1049-001 Lisbon, Portugal

Daniel GARTNER
Cardiff University
School of Mathematics
Cardiff, United Kingdom

Henrique M. G. MARTINS
Centre for Research and Creativity in Informatics (CI2)
Hospital Prof. Doutor Fernando Fonseca
IC-19 Venteira
2720-276 Amadora, Portugal

Corresponding author: José Carlos Ferrão
(jcn.ferrao@gmail.com)

Keywords electronic health record · structured data · resource demand · temporal analysis · data mining · feature selection

Acknowledgements The authors are grateful for the close collaboration with colleagues at the Hospital (BLINDED) and their availability throughout this research study. The authors sincerely thank the associate editor and the anonymous referees for their careful review and excellent suggestions for improving this paper.

Declarations

Funding

BLINDED

Conflicts of interest/Competing interests

No conflicts of interest to declare.

Ethics approval

Ethics approval was waived reviewed and signed off by hospital board and chief information officer. This was a fully retrospective study with anonymized, routinely collected EHR data extracted by a hospital-designated data handler.

Consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

Not applicable

Code availability

Not applicable

Highlights

- This article proposes a data mining framework to inform hospital resource management by predicting managerial variables from electronic health record (EHR) data during the course of inpatient episodes.
- This framework entails a series of systematic procedures for dataset preparation and model building to predict 5 managerial variables: diagnosis categories, procedures, diagnosis-related groups (i.e. episode classification), outlier episodes and patient length of stay.
- Our results reveal that several managerial variables of interest can be predicted from structured EHR data within early stages of inpatient episodes. This can directly inform hospital managers in tactical and operational decision making for resource management.
- The multiple experiments presented in this article provide insights for EHR system developers, data scientists/modelers and system users on how to optimize the value of this methodology, by focusing on certain patient subpopulations and promoting data quality for EHR content areas that, according to results, most impact predictions of managerial variables.

1 Introduction

Cost containment measures such as the US Affordable Care Act [1] and budget constraints put healthcare managers under pressure to efficiently allocate scarce resources. Key challenges of efficient resource management include lack of timely and accurate information [2] and the fact that healthcare resource capacity is perishable [3]. Resource demand prediction can be highly valuable for healthcare managers in making resource allocation decisions and minimizing capacity idle times – this includes aspects such as capacity planning, dynamic bed management, shift scheduling and patient assignment [4]. However, predicting demand and understanding inpatient service needs is challenging as this information is often not available – clinicians might not fully document resource needs as they become known. There are also gaps in resource demand aggregation tools, lack of effective mechanisms to detect changes in patient resource needs and insufficient communication channels to surface relevant information to hospital managers [5].

Electronic health record (EHR) systems have been a cornerstone of health system modernization and potentially contribute to care quality, cost savings and health gains [6, 7]. Routine use of these systems, particularly in hospital settings, produces vast amounts of data even in early stages of inpatient episodes. Also, these data are increasingly recorded in structured formats [8] as opposed to free-text. Valuable insights can be obtained through the secondary use of EHR data [9]. Data mining and predictive modeling are extensive research fields with numerous tools and applications in healthcare [10, 11], which include assisting diagnosis, treatment recommendation and patient readmission prediction [12]. Literature shows the value of these methodologies in informing clinical practice, predicting patient health status and outcomes, particularly in hospital settings [13], these can potentially help identify, integrate and communicate expected demand for hospital inpatient services.

While multiple studies have addressed the use of EHR data for clinical decision support [14], fewer studies have explored how EHR data can be leveraged to inform resource management during the course of inpatient episodes. Also, EHR data provides the basis for diagnosis and procedure coding [15] as well as patient classification systems such as diagnosis-related groups (DRGs) – which establish a linkage between patient clinical characteristics and patterns of resource demand [16]. However, numerous hospitals do not possess mechanisms to provide preemptive predictions of expected resource needs and DRG classification. Instead, episode

coding and classification are only triggered after discharge, and resource use efficiency can only be analyzed retrospectively. Although mechanisms for early DRG classification exist, there is still lack of up-to-date information at each stage of patient care. This requires novel approaches for early prediction [17] and efficient resource allocation decisions such as patient scheduling [18, 19].

With the continuous adoption of EHR systems, there is scope to leverage EHR data during the course of inpatient episodes to provide relevant and timely insights to inform resource allocation decisions. Data mining and predictive modeling methods can be particularly valuable in obtaining managerial insights, which can be updated as more information becomes available.

In this context, this article proposes a methodology to enable systematic use of structured EHR data throughout the course of inpatient episodes and provide key insights to inform managers in making tactical and operational resource allocation decisions. Specifically, this study develops models to predict patient diagnosis categories, procedure codes, DRGs, outlier episodes and LOS. To further refine this methodology, we investigate the influence of population homogeneity on model performance and assess whether tailoring the data mining approach to patient subgroups of (frequently heterogeneous) hospital populations is valuable. We also investigate which EHR data elements have higher influence on model performance.

The remainder of this article is structured as follows: Section 2 reviews relevant literature on the use of data mining and EHR systems, with a focus on managerial insights. Section 3 describes the proposed methodology to obtain predictions of relevant managerial variables during the course of inpatient episodes, to evaluate performance on subpopulations and to assess relevance of EHR data elements to overall model performance. Section 4 presents results from experiments using a dataset of inpatient episodes from a public hospital in Portugal, and section 5 discusses key results, implications and limitations of this study. Section 6 summarizes contributions and outlines future work.

2 Literature review

There has been extensive data mining research in clinical medicine [20]. In this section, we provide a brief literature review focused on prediction of managerial variables related to this study: clinical profiles, procedures, episodes (DRG) classification, LOS outliers and duration.

Prediction of patient clinical profiles has typically focused on using data mining tools to support the di-

agnosis of specific conditions such as liver disease [21], diabetes [22], cancer [23, 24] and cardiac diseases [25]. In these studies, feature sets built based on clinical features are often disease specific and tailored for each case, which makes generalization challenging. A study from Kocbek et al. (2016) [26] uses text mining to predict diagnoses for inpatient episodes from free-text and evaluates the value of different data sources. From a different perspective, prediction of diagnosis classification systems (i.e. clinical coding) has been addressed in numerous studies [15], applying a wide variety of machine learning algorithms to EHR data. The vast majority of these studies use free-text data, which also brings challenges of extracting relevant information from narratives [27]). However, a few studies have approached the use of structured data for coding support and show promising results in overcoming barriers of using free-text EHR data [28](BLINDED)(BLINDED).

Unlike diagnosis coding, procedure coding has been less common in research literature. Chiaravalloti et al. (2014) [29] proposed a system to predict diagnosis and procedure codes using natural language processing and knowledge bases, while Subotin and Davis (2014) [30] developed a system to estimate confidence scores through levels of abstraction. These two studies underline the importance and illustrate the feasibility of predicting procedure codes to mitigate manual workload and provide insights for operational resource management, however none of the studies approached this topic from a temporal perspective during inpatient episodes.

In terms of predicting episode classification, literature is also relatively scarce in spite of general acknowledgement that of its importance for upstream operations scheduling and planning [18, 31]. DRG classification using routinely collected data has been addressed by Gartner et al. (2015) [17] who evaluated different feature selection and classification algorithms on a variety of performance metrics and levels of detail. The temporal classification problem's performance revealed that before admission, at admission and during different stages of care, managers can be provided with more accurate DRG information as compared to a baseline approach, i.e. using a so-called DRG grouper. In addition, related work by Okamoto et al. (2012) [32] addresses prediction of diagnosis-procedure codes (the Japanese equivalent of DRGs) using machine learning models, though not considering the temporal dimension.

Since LOS is a significant proxy of resource consumption, there has been extensive research on LOS prediction through aggregate LOS estimations [33], testing distribution fitting [34] and, more frequently, providing patient-level numerical LOS predictions. Early research started with techniques such as subjective

expert estimation [35] and evolved into risk scores (such as SEWS, APACHE, and SAPS) to explain observed LOS in areas such as stroke [36], cardiac surgery [37, 38] and inpatient or intensive care (ICU) departments [39]. Markov models have also been used to estimate ICU LOS [40]. Data mining models have increasingly played an important role in LOS prediction, either using classification algorithms to predict discrete LOS intervals [41] or regression models to predict numerical LOS values. Although higher performance is reported for classification approaches, these have limitations in that selected LOS thresholds may be arbitrary [42] and not relevant for hospital management. DRG trim points are an example of meaningful LOS thresholds [16]. As such, predicting LOS values through regression is, in principle, more relevant for hospital managers. Multiple regression-based approaches are found in the literature: multilinear regression [43], Poisson regression and negative binomial models [44], generalized linear models and Cox proportional hazards [45]. Data mining regression methods used include neural networks, regression trees, random forests [46, 47] and bagged decision trees [48, 49]. The temporal dimension has been addressed either by assuming which information is available at each stage [50, 51] or by systematically building models using data available at each stage [52, 53].

In summary, there is a lack of holistic approaches to predict multiple managerial variables in a timely manner. This article aims to fill this gap by developing models that can adapt dynamically throughout the course of inpatient episodes and make use of routinely-collected EHR data available in structured formats.

3 Systematic data mining framework

The proposed data-driven methodology uses historical inpatient datasets to model relationships between EHR data and relevant managerial variables (MVs) using supervised learning paradigm to build prediction models. Inpatient episodes are represented as feature sets linked to known outcomes [54]. After training, these models provide predictions on new unseen instances and performance is evaluated using cross-validation. This section describes our methodology to leverage routinely collected EHR data to predict MVs – we explain how MVs are structured from the EHR dataset, describe the predictive modeling methodology and provide an overview of the experimental design of this article.

3.1 Managerial variables and EHR dataset

Each MV represents a relevant resource management element, covering patient clinical profiles, expected resource demand and payment for care services. These MVs are identified as MV1 through MV5 and are structured by combining the healthcare management perspective with data available in the EHR as represented in Figure 1. MVs represent outcome variables used for predictions and are available in administrative (admission-discharge-transfer system) systems as well as in the national database of hospital episodes, which contains diagnosis, procedure and DRG codes. This study was developed using a commercial EHR system (Soarian [55]), a patient-centered system which uses structured data, controlled fields and terminologies, for all essential patient information. For diagnoses specifically, the EHR system had multiple catalogs configured (ICD-9-CM, ICD-10 and working diagnoses) to provide flexibility to clinicians in selecting diagnoses. Table 1 provides details on the contents of each EHR data element.

MV1: Diagnoses Diagnoses include principal and secondary conditions representing clinically meaningful information on patient health status and can provide insights into potential complications and comorbidities. Diagnoses are linked to demand for certain health services (e.g. demand for specific clinical specialists, procedures, materials and overall ward/resource capacity) and are modeled through the International Classification of Diseases (ICD-9-CM version, the coding standard in the Portuguese National Health Service – NHS – at the time of data extraction). Diagnosis codes are available for all historical episodes as coding occurs after discharge and were grouped at the category (3-characters) level to model this outcome variable, to compensate for imbalance (see [56] for an example of typical ICD dataset imbalance) while still providing meaningful insights on clinical profiles. Since multiple diagnosis codes are assigned to each episode, we developed binary classifiers for each ICD category (i.e. each 3-character code) and averaged model performance across all categories considered [57].

MV2: Procedures Procedures entail a wide range of diagnostic exams, procedures and treatment interventions with variable resource intensity and can directly help plan capacity for medical equipment/facilities, skilled human resource needs (clinicians, nurses and technicians), as well as materials (specifically drugs, medical tools and consumables) to support these procedures. Predicting procedures can therefore help plan resource

capacity. Procedures are also modeled using ICD-9-CM codes (described in Volume 3 of this coding scheme), in this case using the full 4-digit procedure codes to capture the differences in complexity and resource needs amongst specific procedures. Similar to MV1, multiple procedure codes are assigned to each episode. We developed binary classifiers for each procedure code and averaged model performance across all codes.

MV3: DRGs Prediction of DRG classification is highly relevant for resource management and planning given its contribution for upstream planning of treatment processes in inpatient episodes [17, 18]. Specifically, DRGs provide insight on the overall clinical profile of the episode, the expected package of medical services and expected payment for hospital episodes. The latter is important for financial forecasting and ensuring hospital resource consumption is managed within expected limits to avoid losses. While DRGs also convey information on clinical profiles, similar to MV1 (albeit at a much more general level), DRGs also convey patterns of resource consumption and expected payment for episodes, which make them a valuable complement to MV1. We used the DRG grouper All-Patients DRG (version 27, the standard in the Portuguese NHS at the time of data extraction). We also modeled DRG prediction as a multi-label classification problem by developing a binary classifier for each DRG code and averaging performance across all DRG codes considered.

MV4 & MV5: Episode outliers and LOS LOS is one of the most widely used proxies of resource consumption and is instrumental for tactical and operational capacity management [18, 58]. Specifically, LOS directly influences capacity for admitting new patients and scheduling elective procedures/surgeries that require bed capacity. Shorter LOS can reduce risks of infections and complications, improve quality of care and promote profit through efficient resource use. However, excessively short LOS might increase risk of readmission and impact quality of care. As such, early LOS prediction is very important for hospital management. We firstly address LOS in terms of outlier episodes (MV4), defined as episodes with LOS outside the corresponding DRG trim points. Prediction of these outliers is relevant for resource management as an indicator of unexpected clinical complications, be related with inefficient resource use or entail higher risk of readmission. This is modeled as 2 binary classifiers (LOS lower or LOS higher than trim points) with value 1 if episodes have LOS outside of DRG trim points. These trim points were obtained from official NHS documentation [59]. We analyzed low and high LOS outliers sep-

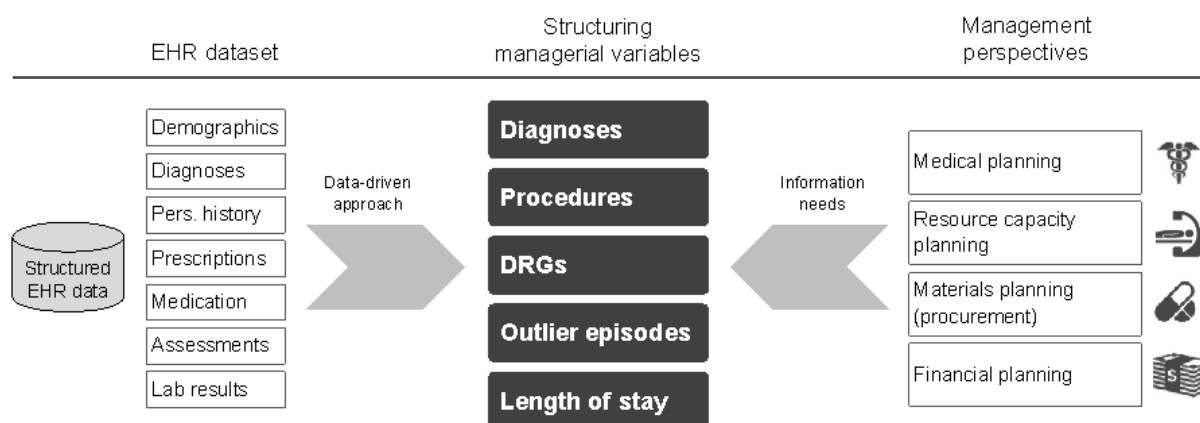


Fig. 1 Conceptual representation of managerial variables, developed from combination of EHR data elements with information needs from managerial perspectives.

arately to potentially surface different factors influencing these two scenarios. LOS (MV5) is in turn modeled as the time (in days) between patient admission and discharge. LOS data was retrieved from the national episode database, in order to ensure alignment with the information reported by the hospital. Since LOS is a numerical value, predictions were obtained with supervised regression methods.

Data element	Description	Feature construction procedure	Classification scheme	Data type	# Features
Demographics	Age and sex information	Define features for age and sex Transform age into equal-width bins (10 years)	–	Numerical (age), binary (sex)	2
Diagnoses*	Working diagnoses (principal and secondary) assigned by clinicians, selected from system-embedded catalogs (pick-lists)	Map all diagnoses into a single catalog (ICD-9-CM) and truncate at category level; Define binary feature for each unique diagnosis category; assign value 1 or 0 based on presence or absence (respectively) in each episode	ICD-9-CM, ICD-10 and local “working diagnoses” catalog	Categorical	613
Personal history	Relevant personal history selected from a small set of health conditions through checkboxes; personal history is preserved across episodes of the same patient	Define binary feature for each unique personal history item; assign value 1 or 0 based on presence or absence (respectively) in each episode	Local catalog	Categorical	36
Allergies	Allergy conditions selected from a system catalog or entered as free-text (normalized during data processing)	Harmonize/map allergies to general allergen designations (e.g. cat, dog, dust, pollen) and active ingredients (not commercial brands) in case of medication allergies; free-text entries may occur and are converted to corresponding allergen/active ingredient designations; review by medical experts is important in case of high variability and high frequency of free-text entries; Define binary feature for each unique allergy item; assign value 1 or 0 based on presence or absence (respectively) in each episode	Local catalog + Free-text designations	Categorical + Free-text	154
Prescriptions	Medical/nursing procedures, diagnostic and imaging exams, laboratory tests	Harmonize procedure designations and remove non-relevant detail; Define binary feature for each unique prescription item; assign value 1 or 0 based on presence or absence (respectively) in each episode	Local catalog	Categorical	1,253
Medication	Prescribed medication, including dosage and administration pathway	Map medication entries to drug active principles leveraging standard (e.g. RxNorm) or EHR-embedded drug catalog, remove detail on dosage and administration pathway; Define binary feature for each unique drug item; assign value 1 or 0 based on presence or absence (respectively) in each episode	Local catalog	Categorical	506
Assessments	Clinical forms configured for data collection in different events such as respiratory assessment, feeding, fluid balance and elimination, scoring scales, medical and nursing consultation notes, admission and discharge forms; composed mainly of structured (checkboxes, dropdowns, buttons and pick-lists); free-text allowed for narrative comments	Identify and merge redundant fields and lists of values across all assessments configured in the EHR system through manual inspection of all assessment fields, supported by system/clinical expert review; Define a feature for each clinically meaningful field, defining feature type (numerical/categorical) in line with underlying clinical concept; Implement algorithm to automatically process EHR database entries and populate data matrix with feature values	–	Numerical, ordinal and categorical;	1,309
Lab results	Laboratory test results, covering mainly analytical essays of biochemistry, hematology, immunology and microbiology.	Harmonize laboratory test labels through manual inspection and supported by expert review; For numerical lab results, define binary features for high and low abnormal flags (if applicable); low/high thresholds should be defined according to limits configured in the EHR system, or if these are not available, they should be defined with clinical experts for categorical results, define categorical feature	Local catalog	Numerical and categorical	1,290

Table 1 Key EHR data elements - description, feature construction procedures, classification schemes and variable types. *Diagnoses represent “working diagnoses” assigned by clinicians during the course of inpatient episodes. These diagnoses are usually temporary and less specific than diagnoses assigned after patient discharge. Therefore, these diagnoses are different from MV1.

3.2 Predictive modeling methodology

Addressing prediction of multiple managerial variables along the course of patient episodes required a systematic data mining approach, by using all structured EHR data available up to a given instant in the episode’s timeline and build prediction models using a standard predictive modeling pipeline (PMP) of data mining tasks. Specifically, this pipeline included EHR dataset construction, feature set construction, feature selection, prediction model development and model evaluation. The key definitions in this context are:

Episode: Represents an instance in our EHR datasets, defined by the period between patient admission and discharge; for each episode e_i contained in the episode set E , there is a collection of database records comprising all data recorded during patient stay.

Feature: A component x_j of the feature space F , used as an independent variable to be used as input for prediction models. Features represent relevant characteristics of inpatient episodes derived from EHR data.

EHR data element: Each block of EHR information b_l from the set of all data elements B ; demographic data, personal history, diagnoses and prescriptions are examples of data elements.

Full EHR dataset: Corresponds to the global collection D of EHR data and includes all database records (tuples) r_m (in raw format) recorded for each episode.

Instant: A given point in the episode timeline from the global set of instants T . Each instant t_k uses a relative time reference expressed in terms of time elapsed after patient admission.

Filtered EHR dataset: A subset D_k of database records r_{mk} (in raw format) of the full EHR dataset D , obtained by retrieving only EHR entries produced between admission and a given instant t_k . Filtered datasets are cumulative in that each filtered dataset is an extension of the filtered dataset constructed for the previous instant.

Data matrix: Tabular format of EHR data used as input to develop prediction models; matrix cells d_{ij} represent the value of feature f_j for episode e_i .

The proposed systematic methodology is implemented by executing standardized PMPs for each instant and each managerial variable of interest, as represented in Figure 2. Each PMP uses a filtered EHR dataset with tuples containing the field labels, corresponding values and date-time stamps (see section 3.1 for details on EHR dataset).

Key activities comprised in PMPs are as follows:

1. *EHR dataset construction:* Creation of filtered EHR datasets from the full EHR dataset for each instant of interest, using date-time stamps of EHR database records. This process was automated using MySQL stored procedures (Figure 3), firstly defining all instants t_k composing the instant set T , then retrieving episode admission times (ta_i) to produce a matrix of temporal boundaries S whose cells $s_{ik} = ta_i + t_k$ represent the (absolute) date and time to filter data through comparison with date-time stamps in the EHR dataset. Time instants were defined through consultation with hospital stakeholders. The density of time instants is higher in the first 24h after admission, during which most patients were admitted in the emergency department and where key decisions of inpatient bed and resource management are made [61]. No limits were placed on maximum or minimum number of datasets – we constructed 14 filtered EHR datasets aligned with the 14 time points defined.
2. *Feature set construction:* Data preprocessing activities to define the feature space to represent inpatient episodes, based on EHR data contents. Table 1 outlines the procedures to construct features from each EHR data element, which included resolving redundancies and data harmonization. Certain features may have missing values, in which case these can be handled with missing value imputation if appropriate, or features might be excluded from analysis if missing rates are excessively high, generally if higher than 50%.
3. *Feature selection:* Consists in the use of methods to define informative subsets of features [62] and mitigate negative effects (e.g. overfitting and poor model interpretability [63]) of high dimensionality, particularly due to a high number of binary features resulting from structured EHR data. Feature selection is key for successful data mining applications [64]. Filter methods were used for classification problems (MV1 – MV4) due to scalability and faster execution times [63]. Specifically, we used fast-correlation-based filter (FCBF) [65], minimal-redundancy maximal-relevance (mRMR) [66] and chi-square [67]. For regression (MV5 – LOS), we used embedded methods which are part of regression models, specifically forward-backward stepwise selection for multilinear regression (starting from empty feature sets, maximum inclusion and minimum removal p-value thresholds set to 0.05 and 0.1, respectively, with infinite number of iterations) [68] and out-of-bag feature importance for bagged decision trees (with ensemble size to 10 trees based on

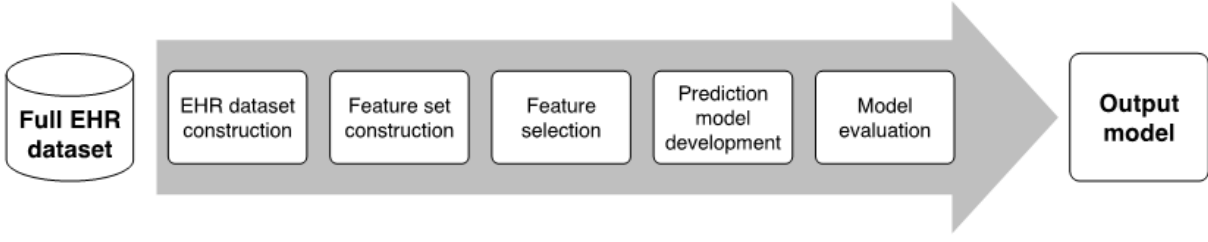


Fig. 2 Activities comprised in a predictive modeling pipeline (PMP) to build prediction models from raw EHR data (adapted from [60]).

preliminary analysis, minimum 5 instances per leaf, sampling fraction of 100% and evaluated all features as candidate splitting criterion). Feature selection is applied for each MV separately, generating MV-specific feature subsets that are then used for model building. All EHR data elements were considered as candidate features for all MVs when applying feature selection methods.

4. *Prediction model development*: Consists in training supervised learning models to predict each managerial variable at each instant of interest. For classification (MV1 – MV4) we used decision trees (CART algorithm in MATLAB (*classregtree*), Gini index as splitting criterion [69], minimum of 2 instances per leaf and post-pruning optimized based on F1-score) and logistic regression models [70] (with maximum likelihood estimation [71], generalized linear models with a sigmoid link function and classification cutoff threshold optimized for each model, by testing all cutoff values between 0 and 1 in steps of 0.005 and selecting the cutoff point with highest F1-score).. For regression (MV5) we applied multilinear regression with forward-backward stepwise selection [68] and bagged decision trees (valuable for unstable models and showing promising results in [49]) using the MATLAB algorithm *TreeBagger* [72, 73].

5. *Model evaluation*: Assessment of model predictive power through cross-validation [74], comparing predictions with ground-truth outcomes in the test set. Performance metrics were obtained across N test instances, based on confusion matrix counts for classification (TP – true positive, TN – true negatives, FP – false positives, FN – false negatives) and predicted (\hat{y}_i) and known (y_i) numerical responses for regression. For MV1, MV2 and MV3 where one classification model is developed for each label considered, model performance metrics were averaged across all labels. These performance metrics were selected in line with metrics typically reported in related literature and are calculated as follows:

Accuracy – A

$$A = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Precision – P

$$P = \frac{TP}{TP + FP} \quad (2)$$

Recall – R

$$R = \frac{TP}{TP + FN} \quad (3)$$

F1-score

$$F1 = \frac{2PR}{P + R} \quad (4)$$

Root mean-squared error – RMSE

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (5)$$

Mean absolute error – MAE

$$MAE = \frac{1}{N} \sum_{i=1}^N \text{abs}(\hat{y}_i - y_i) \quad (6)$$

Mean absolute percentage error – MAPE

$$MAPE = \frac{1}{N} \sum_{i=1}^N \frac{\text{abs}(\hat{y}_i - y_i)}{y_i} \quad (7)$$

Coefficient of determination – R^2

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (8)$$

AUC/ROC metrics were not deemed adequate for highly-imbalanced datasets (i.e. high proportion of negative examples in diagnoses, procedure and DRG prediction) as these typically provide an overly-optimistic view of model performance and may not be meaningful in terms of clinical/operational utility. We instead focused on metrics that focus on TP, FP and FN counts,

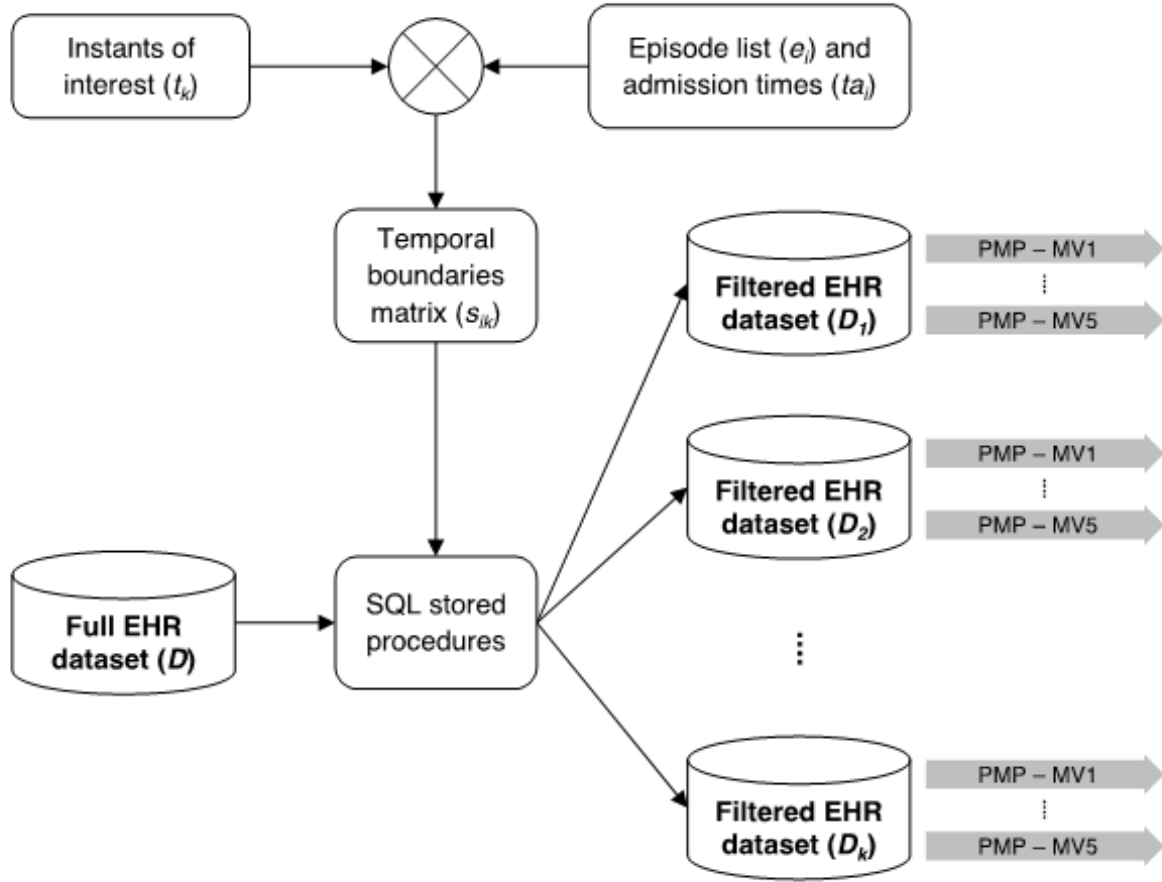


Fig. 3 Mechanism to create filtered EHR datasets for all instants of interest by extracting data from the full EHR dataset using SQL stored procedures and a temporal boundaries matrix.

and prioritized F1-scores as the most important metric to assess performance of classification models.

All feature selection methods and prediction models were executed in MATLAB (Statistics and FEAST Toolboxes). For FCBF and mRMR filter methods [75] we used a 10^{-6} threshold for symmetrical uncertainty to ensure retrieval of all relevant and non-redundant features for each label. Classification models were built in a stepwise forward approach, adding features by order of relevance (as ranked by each filter method). Regression models were developed using log-transformed LOS values (without offset) to compensate for high positive skewness of LOS distribution. All evaluation metrics were calculated using 5-fold cross validation, comparing predicted with actual values.

3.3 Experimental design

In this article, we designed experiments to evaluate the proposed methodology in several perspectives and

generate relevant insights for hospital managers. The methodology starts with problem scoping to define relevant managerial variables MV1 to MV5. The experiments conducted in this article are displayed in Figure 4 and leverage PMPs described above to address different research questions:

- Determine best-performing combination of feature selection methods and prediction models for each managerial variable, by testing these combinations on the full EHR dataset, averaging model performance for multi-label variables MV1-MV3 and selecting the combination with the highest average performance for MV (results presented in section 4.2).
- Conduct a temporal analysis of model performance starting after admission and at specific instances of patient stay; this analysis establishes the base case which serves as comparator for subsequent analyses performed on subpopulations and to assess value of information (results presented in section 4.3).

- Replicate temporal analysis for subpopulations, specifically the two most frequent major diagnostic categories (MDC) in the dataset – respiratory and cardiovascular diseases; as well as extend LOS analysis on age, sex and LOS duration (lower/higher than median LOS) subpopulations; this experiment also includes predicting subpopulation membership, i.e. predict which subpopulation the episode will fall into, to assess whether episodes can be assigned to subpopulations in order to develop more tailored prediction models (results presented in section 4.4).
- Evaluate the contribution of each EHR data element to overall model performance, replicating the base case analysis with removal of one EHR element at a time (demographics, diagnoses, personal history and allergies, prescriptions, medication, assessments and laboratory values and evaluating differences in model performance relative to the base case; results presented in section 4.5).

4 Results

This section presents the results from analyses conducted with a real-world dataset from a large public hospital (approx. 700 beds) in Portugal, including descriptive statistics and model predictive power obtained across the different experiments. Each of the following subsections presents results of specific experiments outlined in the experimental design.

4.1 Dataset overview

The case-study dataset consists of EHR source files of 5,089 non-surgical inpatient episodes collected over a 6-month time span and covering multiple medical specialties (mainly internal medicine, pneumology, infectiology, gastroenterology and nephrology). From the total feature set constructed from the EHR system, 201 features had missing values, all with missing rates over 50%. We excluded these 201 features due to their high missing rates. Table 2 presents the number of features in the dataset after preprocessing.

Data on managerial variables (ICD-9-CM diagnosis and procedure codes, DRG class and corresponding MDC, admission and discharge dates/times and LOS values) were extracted from the Portuguese national DRG database (WebGDH) in which public hospitals report inpatient episodes. Additional information on LOS trim points (to label outlier episodes) was extracted from official NHS documents. Table 3 presents summary statistics of the five managerial variables, which

Data element	Summary Statistics
Demographics	Age: mean = 67.7; median = 72; SD = 18.0; IQR = 26 Sex: 49.5% M; 50.5% F
Data element	Average number of entries per episode
Diagnoses	4.0
Personal history	32.1
Allergies	0.2
Prescriptions	34.6
Medication	12.9
Assessments	59.6
Laboratory results	16.6

Table 2 Feature set overview for the full EHR dataset. SD = standard deviation; IQR = interquartile range.

shows the general imbalance of categorical variables (MV1 – MV4) and a skewed LOS distribution. The box plot depicted in Figure 5 shows the distribution of LOS for the global population as well as for subpopulations segmented by MDC, age, and sex. MDC 4 (respiratory) is shifted towards higher LOS values and MDC5 has a smaller interquartile range with respect to the global population. Additionally, patients with age lower or equal to the median (72 years) exhibited lower median LOS and interquartile range than patients older than 72 years. All subpopulations exhibited a considerable number of outliers with LOS higher than the 99th percentile.

4.2 Full EHR dataset analysis

Results obtained with the full EHR dataset are presented in Table 4 (averaged across labels for classification) and show that the combination of mRMR feature selection and logistic regression models consistently exhibited the highest F1-score, while bagged regression trees outperformed multilinear regression in almost every metric (except for MAPE). For classification models, we prioritized results based on F1-score since this measure combines precision and recall and is typically the most important measure in binary classification problems. For regression models, we prioritized MAE as most meaningful metric as it provides the most direct and clear linkage to LOS values, expressed in days. Based on these results, we selected logistic regression combined with mRMR feature selection and bagged regression trees for subsequent experiments.

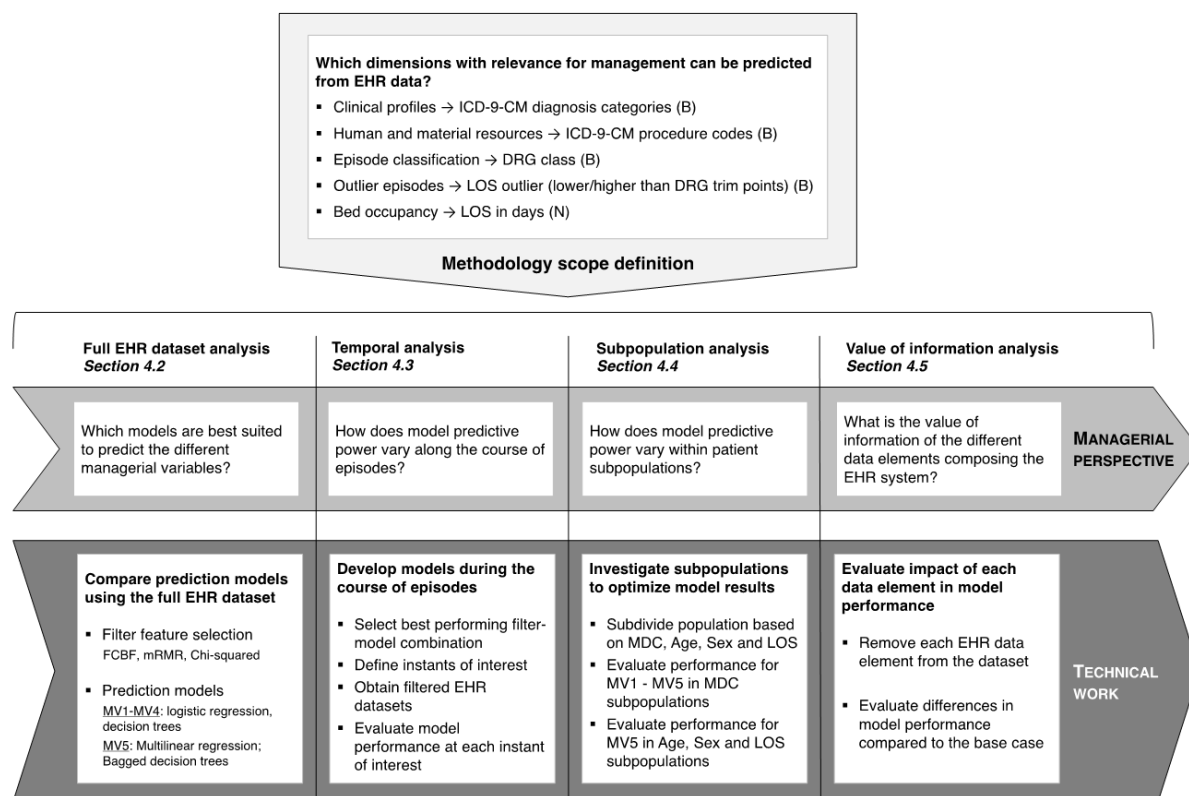


Fig. 4 Schematic representation of the methodological workflow and experimental design. (B) indicates binary managerial variables and (N) indicates a numerical managerial variable.

Managerial variable	Scope	Selected statistics
MV1 Diagnoses	Top 75 ICD-9-CM categories	401 – Essential hypertension (N = 2044) 276 – Disorders of fluid, electrolyte & acid-base balance (N = 1706) 250 – Diabetes mellitus (N = 1395) V58 – Encounter for other/unspec. procedures & aftercare (N = 1313) 272 – Disorders of lipid metabolism (N = 1258)
MV2 Procedures	Top 75 ICD-9-CM procedures	90.59 – Microscopic examination of blood NEC (N = 4710) 87.44 – Routine chest X-ray (N = 3398) 89.52 – Electrocardiogram (N = 3150) 99.18 – Injection/infusion of electrolytes (N = 2764) 89.65 – Arterial blood gas measure (N = 2375)
MV3 DRGs	Top 25 AP-DRG codes	541 – Simple pneumonia & other respiratory disorders excl. bronchitis, asthma (N = 379) 89 – Simple pneumonia and pleurisy, age >17, with CC (N = 335) 127 – Heart failure and shock (N = 230) 544 – Congestive heart failure and cardiac arrhythmia, major CC (N = 180) 410 – Chemotherapy (N = 144)
MV4 LOS outliers	Low/High LOS outliers	Low outliers – N = 428 High outliers – N = 168
MV5 LOS	LOS in days (integer)	Mean = 9.03; Median = 7; SD = 9.86; IQR = 10

Table 3 Scope and summary statistics of managerial variables MV1 to MV5. The top 5 most frequent categories are presented for MV1, MV2 and MV3. NEC = not elsewhere classified; CC = complications or comorbidities. SD = standard deviation; IQR = interquartile range.

(a)

MV	Precision						Recall					
	DT			Logit			DT			Logit		
	FCBF	mRMR	Chi	FCBF	mRMR	Chi	FCBF	mRMR	Chi	FCBF	mRMR	Chi
Diag.	0.727	0.741	0.739	0.656	0.649	0.645	0.468	0.489	0.481	0.590	0.599	0.586
Proc.	0.671	0.659	0.644	0.559	0.583	0.537	0.474	0.507	0.490	0.670	0.685	0.681
DRG	0.329	0.448	0.401	0.369	0.474	0.378	0.186	0.279	0.241	0.567	0.553	0.549
LOS	0.563	0.508	0.510	0.364	0.398	0.359	0.245	0.319	0.220	0.594	0.579	0.530
out.												

MV	F1-score						Accuracy					
	DT			Logit			DT			Logit		
	FCBF	mRMR	Chi	FCBF	mRMR	Chi	FCBF	mRMR	Chi	FCBF	mRMR	Chi
Diag.	0.550	0.569	0.562	0.608	0.611	0.598	0.952	0.953	0.952	0.946	0.946	0.944
Proc.	0.487	0.529	0.508	0.590	0.615	0.580	0.946	0.948	0.946	0.935	0.942	0.935
DRG	0.217	0.328	0.289	0.423	0.481	0.416	0.979	0.980	0.978	0.966	0.974	0.967
LOS	0.338	0.385	0.296	0.437	0.467	0.403	0.944	0.945	0.943	0.903	0.926	0.902
out.												

(b)

MV	RMSE		MAE		MAPE		R ²	
	MLR	BT	MLR	BT	MLR	BT	MLR	BT
LOS	9.816	5.912	3.787	3.168	0.404	0.474	-0.015	0.646

Table 4 Average model performance obtained through 5-fold cross validation for (a) classification (decisions trees – DT – and logistic regression - Logit) and (b) regression (multilinear regression – MLR – and bagged regression trees – BT) models using the full EHR dataset, combined with feature selection methods fast correlation-based filter (FCBF), minimal-redundancy maximal-relevance (mRMR) and chi-squared (Chi). The best results obtained for each managerial variable (MV) in each metric (RMSE – root mean squared error, MAE – mean absolute error, MAPE – mean absolute percentage error and R² – coefficient of determination) are presented in boldface.

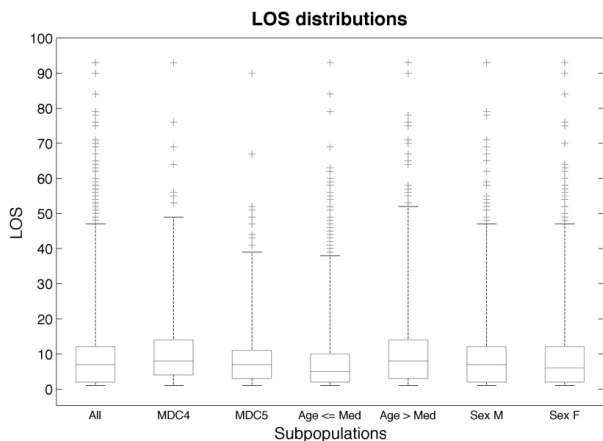


Fig. 5 Box plot distributions of LOS on the global inpatient population and on subpopulations segmented by MDC, age, and sex. Outliers (crosses) represent episodes with LOS higher than the 99th percentile.

4.3 Temporal analysis

Next, we analyzed the evolution of predictive power along the course of episodes. 14 instants of interest were defined between 1 hour and 5 days after admission (roughly up to half the average LOS), as represented in Figure 6, for which 14 filtered EHR datasets were con-

structed. These filtered datasets were processed using a stored procedure (see Figure 3) and a MATLAB algorithm to read source files, perform data preprocessing and populate data matrices (BLINDED).

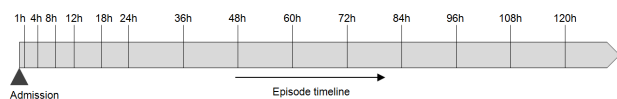


Fig. 6 Relevant time points for the temporal analysis.

Figure 7 depicts the evolution of model predictive power within 5 days after admission, exhibiting comparable patterns in charts (a) to (d) with performance starting off at lower levels and increasing steeply in the first 20 to 40 hours. Average F1-scores were low for high LOS outliers (below 30%) and for DRGs (below 50%) than for diagnoses and procedure (both above 50%). For LOS regression models, a significant decrease in error-based metrics, especially for MAPE which starts at values over 200% and drops to values under 60% after 5 days of stay. R² also improves steadily during the time frame. For label-level detail, Table 5 presents the top 5 and bottom 5 performing labels and shows high performance variability which

is not captured in average measures. The five best diagnosis and procedure labels achieve high F1-score values; DRGs achieve slightly lower results. Some of these top performing cases show good performance at early episode stages. A list with descriptions of diagnoses, procedures and DRG codes is presented in the appendix.

Overall, these results show that performance improves as more data become available for prediction models during the course of episodes. The increasing availability of data is enabled by the use of filtered EHR datasets - the higher the corresponding time instant, the more data these filtered EHR datasets contain (i.e., EHR data collected for that episode from admission and up to a certain amount of hours after admission). No episodes were excluded from the temporal analysis when their LOS is lower than the instant being considered - in such cases, filtered EHR datasets include all data collected for those episodes, and sample sizes remain constant throughout the temporal analysis.

To obtain further insights on the extent to which the average results in Table 4 have practical managerial utility, we observed (Table 6) that diagnoses and procedures have higher proportion of labels reaching F1-scores above 50%, while for DRGs this only happens for 9 out of 25 DRGs in scope. The 5 DRGs with higher reimbursement rates (Table 7) show that this methodology produces acceptable predictive performance for DRG 14 (stroke with infarction) with F1-scores higher than 50% at 12 hours after admission (below expected LOS for this DRG) and for DRG 557 (hepatobiliary and pancreas disorder, with comorbidities/complications), where models surpassed 50% F1-score at 72 hours after admission. For LOS regression, the scatter plots in Figure 8 show performance improvement from charts (a) to (d) (with points shifting towards the diagonal). Models seem to underestimate LOS given the higher concentration of points below the diagonal. The breakdown analysis with performance measures computed for different LOS quartiles in Table 8 shows that RMSE and MAE increase for subgroups with higher LOS, whereas MAPE decreases considerably for episodes with higher LOS.

4.4 Subpopulation analysis

Subpopulation analysis included model evaluation for all managerial variables in clinical subgroups MDC4 and MDC5, as well as outlier episodes and LOS values for age (lower or equal to the median; and higher than the median), sex and LOS (higher than 1; lower

or equal to the median; and higher than the median) subgroups.

MDC subpopulation results in comparison with the base case (section 4.3) indicate that increasing training set homogeneity does not influence results consistently for diagnoses, procedures or DRG labels (see Table 9), while it seems to improve performance of LOS outliers. LOS prediction worsened for MDC4 but improved slightly for MDC5. Observing F1-score differences higher than 10% in comparison with the base case (Table 10), we find clinical correlation between these labels and the corresponding MDC. This points towards potential benefit in using homogeneous populations to train prediction models. Several of these labels are relevant in terms of patient health status (e.g. need for cardiopulmonary resuscitation and mechanical ventilation) and of hospital resources (e.g. CT scan, bronchoscopy and thoracentesis).

Additional subpopulation analyses performed for outlier episodes and LOS (presented in Table 11) showed performance variations for both age and LOS subpopulations, while male/female subpopulations did not show significant change. High LOS outliers and LOS regression models performed better for patients younger than 72 years old, while models for low LOS outliers decreased performance for younger patients. For LOS subpopulations, the main variations are observed in lower performance for low LOS outliers and LOS in $LOS > 1$ and $LOS > \text{median}$ subpopulations; and LOS regression models improve performance for patients with $LOS \leq \text{median}$ subpopulations (significant improvement in MAE relative to the base case).

While model performance improves for subpopulations, the development of subpopulation-specific models impacts their generalizability, requires additional computational effort to build models for multiple subpopulations, and may introduce biased predictions if incorrect subpopulations are used, which introduces important trade-offs. Having models that perform better on subgroups is only relevant and useful if we could know beforehand which MDC the episode belongs to, so that model development can already be targeted for that subpopulation. In order to assess feasibility of determining MDC beforehand (since this information is not directly available in the EHR), we also developed models to predict subpopulation membership, i.e. predict which subpopulation the episode will fall into. A standard PMP with mRMR and logistic regression was applied on filtered EHR datasets to predict subpopulation membership. These results (Table 12) indicate that it is possible to predict subpopulation membership with confidence from early stages of patient stay, especially for LOS subpopulations.

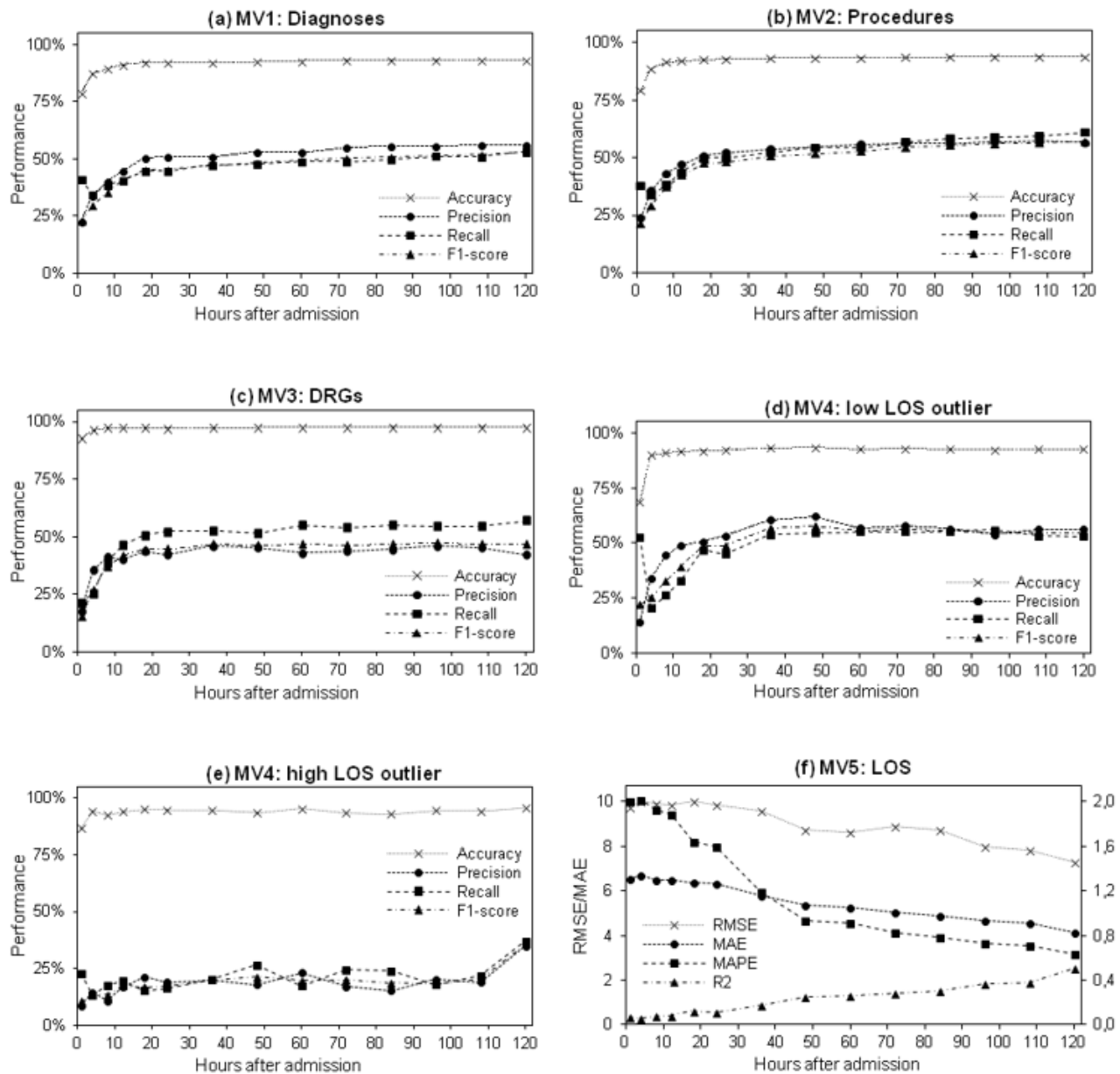


Fig. 7 Performance of prediction models developed within the first 120 hours of inpatient episodes for (a) diagnoses, (b) procedures, (c) DRGs, (d) low LOS outliers, (e) high LOS outliers and (f) LOS (for the latter, MAPE and R^2 are represented on the right axis).

4.5 Value of information analysis

The outputs of this experiment are represented as symbols in Table 13, showing intensity of influence of removal of EHR data elements at 120h after admission. Personal history and allergies were here considered as one EHR element since both represent chronic conditions. Results show that all data elements exert positive influence on model performance, i.e. the removal of these elements resulted in lower model performance. Overall, demographics, personal history and laboratory

results marginal influence in overall performance, while prescriptions had a medium-strong influence on models predicting procedures during the episode. Conversely, diagnoses (assigned by clinicians during episodes), medication and assessment data had a strong influence in one or more managerial variables. Specifically, clinician-assigned diagnoses highly impacted ICD-9-CM diagnosis and DRG prediction; medication and assessment data impacted both outlier episodes and LOS prediction.

(a)								
Diagnoses			Procedures			DRG		
Category	1h	120h	Code	1h	120h	Code	1h	120h
714	0.575	0.914	90.59	0.962	0.963	410	0.672	0.777
434	0.164	0.878	87.44	0.815	0.848	450	0.399	0.732
E950	0.558	0.868	54.91	0.214	0.845	139	0.110	0.672
345	0.153	0.846	39.95	0.180	0.843	14	0.084	0.653
427	0.476	0.812	99.28	0.704	0.824	202	0.190	0.609

(b)								
Diagnoses			Procedures			DRG		
Category	1h	120h	Code	1h	120h	Code	1h	120h
459	0.107	0.243	99.03	0.014	0.198	87	0.136	0.205
426	0.149	0.266	33.24	0.015	0.279	316	0.092	0.301
V43	0.143	0.281	87.42	0.065	0.292	90	0.111	0.323
V10	0.143	0.287	90.52	0.102	0.301	88	0.099	0.331
V87	0.133	0.299	91.32	0.127	0.305	566	0.060	0.363

Table 5 F1-scores obtained for individual labels of diagnoses, procedures and DRGs. (a) Top 5 labels with best performance at 120h; (b) Bottom 5 labels with worst performance at 120h.

Diagnoses							
	1h	12h	24h	48h	72h	96h	120h
$F1 > 50\%$	4	19	29	34	36	38	38
$F1 > 70\%$	0	6	7	10	11	11	12
$F1 > 80\%$	0	2	3	4	4	5	5

Procedures							
	1h	12h	24h	48h	72h	96h	120h
$F1 > 50\%$	8	9	16	19	27	29	35
$F1 > 70\%$	5	5	6	7	7	7	12
$F1 > 80\%$	2	3	3	3	3	3	4

DRGs							
	1h	12h	24h	48h	72h	96h	120h
$F1 > 50\%$	1	2	3	7	8	8	9
$F1 > 70\%$	0	1	1	2	2	2	2
$F1 > 80\%$	0	0	0	0	0	0	1

Table 6 Number of labels with F1-score higher than 50%, 70% and 80% for diagnoses, procedures and DRGs (MV1 to MV3), at different instants.

5 Discussion

5.1 Analysis of results

The full EHR dataset experiment showed that logistic regression models outperformed decision trees in F1-score, accuracy and recall, yielding lower false negative rates. This is relevant for managerial variables such as diagnoses and procedures for which false negatives might represent gaps in clinical documentation and procedure reporting/billing. Similarly, low occurrence of false negatives is relevant to timely identify DRG clas-

sifications and outlier episodes to draw attention to potential deviations in health status and treatments processes. In feature selection methods, we observed that chi-square was always outperformed by FCBF and mRMR methods, showing that multivariate feature selection is more suitable.

Model performance values obtained in this study are comparable to related literature, being lower than [76, 77], higher than [56] and similar to [78, 79] as well as other NLP-based studies. For DRG prediction, our results are comparable to Gartner et al. (2015) [17] (comparison of accuracy is difficult since authors used a multi-class model). We focused on F1-score metrics as these are more realistic in terms of practical applicability, especially in highly imbalanced datasets where accuracy tends to be over-optimistic. For LOS prediction, bagged regression trees outperformed multilinear regression in RMSE, MAE and R^2 metrics, providing results with better practical application overall. Our results are comparable with Verburg et al. (2014) [45] and lower than Xie et al. (2015) [48]. While these comparisons must always consider inherent variability in problems and dataset complexity, our results are in line with related literature.

The base-case temporal analysis showed similar performance patterns across managerial variables, with steeper increase in the first 1-2 days. Average performance reaches acceptable values (above 50%) for diagnoses and procedures, but stays slightly below this threshold for DRGs. For DRGs, this might be related with very low class representation but also lack of alignment between EHR information for care provision and DRG classification purposes. In effect, primary and sec-

DRG	Value	1h	12h	24h	48h	72h	96h	120h
533	7.219,98 €	0.076	0.356	0.361	0.401	0.429	0.446	0.476
14	4.804,34 €	0.084	0.579	0.631	0.632	0.663	0.657	0.653
557	4.683,06 €	0.107	0.423	0.469	0.483	0.503	0.533	0.504
552	4.227,63 €	0.084	0.309	0.358	0.447	0.379	0.433	0.401
15	3.924,43 €	0.050	0.302	0.378	0.422	0.422	0.428	0.432

Table 7 Temporal results of F1-scores for the 5 DRGs with highest reimbursement rates (according to NHS reimbursement rates) [59].

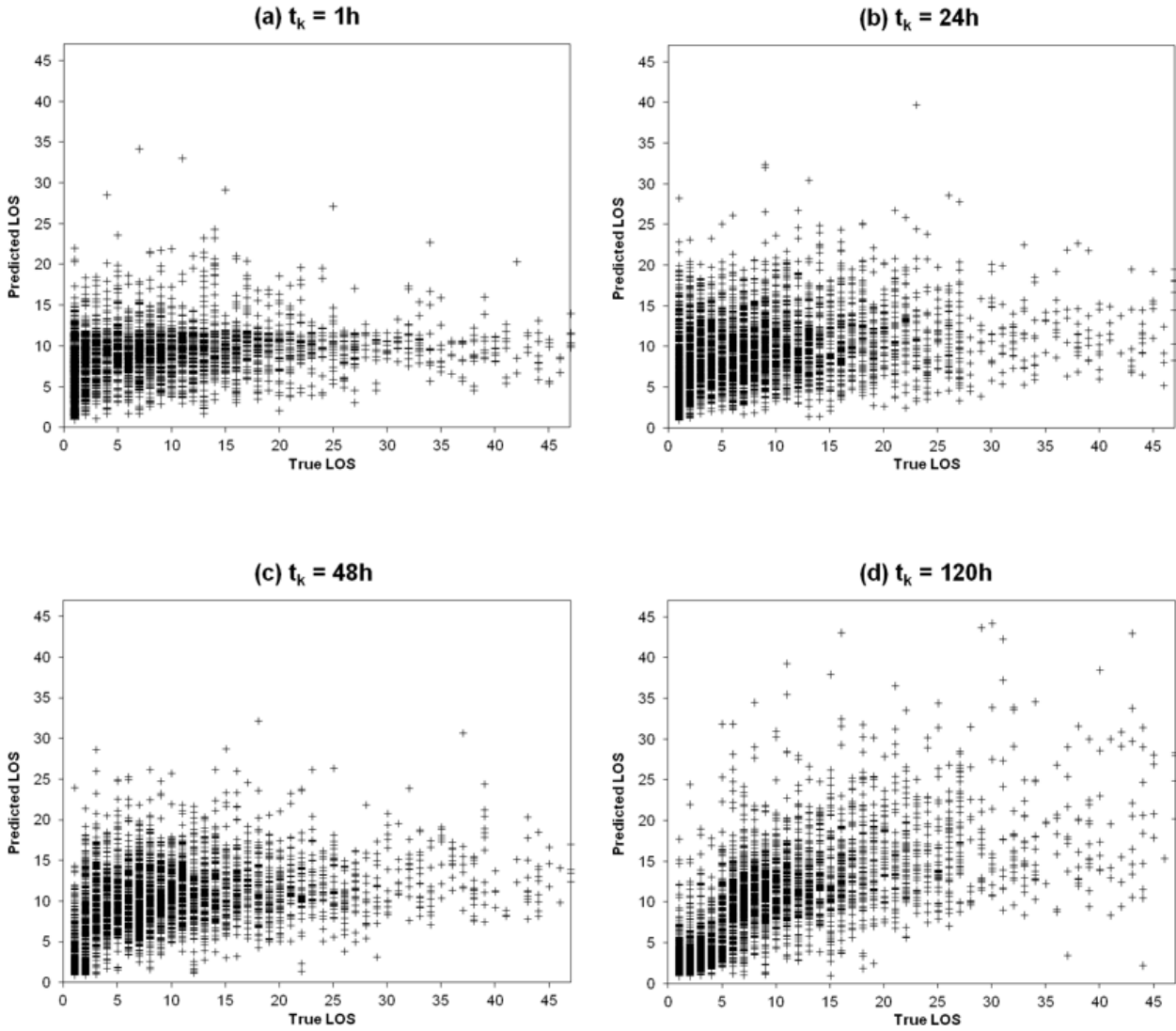


Fig. 8 Scatter plots with predicted vs. ground truth LOS values, obtained with bagged regression models at 1h, 24h, 48h and 120h after admission.

LOS subgroup	RMSE	MAE	MAPE
LOS = 1	1.938	1.074	1.074
2 ≤ LOS < 7	3.773	2.398	0.653
7 ≤ LOS < 12	4.757	3.593	0.430
LOS ≥ 12	12.126	7.993	0.328

Table 8 Breakdown analysis of bagged regression tree performance at 120h after admission, by LOS groups.

ondary diagnoses very often differ from clinical and billing perspectives. Low class representation may also have impacted performance for high LOS outliers, although these can be associated with numerous factors (potentially only discovered at later stages of episodes). Generally, logistic regression provided high confidence predictions for multiple labels (especially within diag-

Diagnoses		1h	12h	24h	48h	72h	96h	120h
Base case		0.229	0.408	0.454	0.485	0.504	0.517	0.531
MDC4		0.214	0.409	0.453	0.486	0.493	0.511	0.523
MDC5		0.229	0.410	0.446	0.471	0.490	0.502	0.531
Procedures		1h	12h	24h	48h	72h	96h	120h
Base case		0.212	0.423	0.481	0.518	0.545	0.562	0.570
MDC4		0.188	0.355	0.404	0.435	0.448	0.487	0.500
MDC5		0.171	0.317	0.392	0.404	0.444	0.448	0.481
DRGs		1h	12h	24h	48h	72h	96h	120h
Base case		0.132	0.434	0.479	0.503	0.496	0.504	0.497
MDC4*								
MDC4		0.258	0.405	0.442	0.476	0.485	0.487	0.495
Base case		0.339	0.589	0.610	0.624	0.621	0.627	0.637
MDC5*								
MDC5		0.331	0.559	0.609	0.625	0.626	0.639	0.650
LOS outliers		1h	12h	24h	48h	72h	96h	120h
Base case low outliers		0.219	0.391	0.486	0.579	0.563	0.547	0.543
MDC4		0.271	0.501	0.634	0.738	0.704	0.685	0.669
MDC5		0.167	0.396	0.519	0.679	0.676	0.670	0.695
Base case high outliers		0.104	0.173	0.174	0.211	0.199	0.183	0.356
MDC4		0.119	0.258	0.330	0.333	0.389	0.316	0.447
MDC5		0.133	0.292	0.315	0.401	0.379	0.375	0.665
LOS		1h	12h	24h	48h	72h	96h	120h
Base case		6.353	6.297	6.142	5.186	4.863	4.481	3.964
MDC4		6.903	6.865	6.206	5.894	5.719	4.989	4.533
MDC5		5.576	6.090	5.599	4.954	4.964	4.438	3.766

Table 9 Comparison of model performance obtained in Major Diagnostic Categories 4 and 5 (MDC4 – Respiratory system and MDC5 – circulatory system) subpopulations with the original results obtained for the base case (overall population). Results are expressed in terms of F1-scores for diagnoses, procedures, DRGs and LOS outliers, and as MAE for LOS. The asterisk (*) denotes that DRG prediction models were built from the MDC4 and MDC5 subpopulations only for DRGs existing in the corresponding MDC. For comparison with MDC subpopulation results, the DRG base-case was calculated using only the performance for DRGs of each MDC. *Not applicable, since very few (or zero) episodes exist in these categories.

noses and procedures), which indicate that these models can anticipate clinical profiles and resource needs even at earlier episode stages, as well as identify episodes with high resource intensity and reimbursement rates. This information might be most valuable if preemptively provided to managers for tactical and operational resource planning.

LOS prediction models reveals steadily decreasing error rates, with acceptable MAE and MAPE values when approaching 5 days after admission (error values are comparable to Huang et al. (2013) [52]). Generally, bagged decision trees tend to underestimate longer-stay episodes (also evidenced in Table 8). This behavior

might be explained by the highly skewed distribution – high prevalence of low LOS episodes decreases predictions in leaf nodes and leads to systematic underestimation. The breakdown analysis shows that these models are able to provide valuable predictions for up to 75% of inpatient episodes with $LOS \leq 12$ days.

In subpopulation analysis, while performance did not improve significantly overall, improved results in specific cases/categories (diagnoses, procedures, LOS) can contribute to the value of this methodology in real-world settings, notably in LOS prediction for younger patients and shorter stay ($LOS \leq 8$ days) with under 1 day error. This also corroborates findings from the

MDC4 Diagnoses	Procedures
511 Pleurisy	33.24 Closed bronchial biopsy
438 Late effects of cerebrovascular disease	99.60 Cardiopulmonary resuscitation NOS
493 Asthma	99.03 Whole blood transfusion NEC
593 Other disorders of kidney and ureter	33.22 Fiber-optic bronchoscopy
491 Chronic bronchitis	88.38 Other CT scan
87.44 Routine chest x-ray	
93.90 Non-invasive mech ventilation	
33.24 Closed bronchial biopsy	

MDC5 Diagnoses	Procedures
426 Conduction disorders	88.38 Other CT scan
459 Other disorders of circulatory system	89.50 Ambulatory cardiac monitoring
428 Heart failure	88.77 Diagnostic ultrasound-vascular
424 Other diseases of endocardium	89.52 Electrocardiogram
412 Old myocardial infarction	96.71 Continuous invasive mech ventilation <96h
402 Hypertensive heart disease	34.91 Thoracentesis

Table 10 Diagnosis and procedure labels (ICD-9-CM) exhibiting an improvement in F1-score higher than 10% (with respect to the base case) in MDC4 (respiratory) and MDC5 (circulatory) subpopulations. NOS = not otherwise specified. NEC = not elsewhere classified.

	Low LOS outliers	High LOS outliers	LOS
Base case	0.543	0.356	3.964
Age \leq median	0.513	0.393	3.513
Age $>$ median	0.600	0.433	4.456
Male	0.561	0.348	4.015
Female	0.554	0.351	4.040
LOS $>$ 1	0.383	0.389	4.671
LOS \leq median	0.542	NA*	0.938
LOS $>$ median	NA*	0.370	7.280

Table 11 Comparison of model performance at 5 days (120h) after admission for age, sex and LOS subpopulations, obtained with logistic regression (F1-score performance) and bagged regression trees (MAE performance) models for LOS outliers and LOS, respectively.

	1h	12	24h	48h	72h	96h	120h
MDC 4	0.444	0.663	0.726	0.738	0.739	0.740	0.728
MDC 5	0.363	0.650	0.683	0.678	0.675	0.680	0.675
LOS > 1	0.905	0.923	0.948	0.953	0.936	0.943	0.947
LOS \leq Med	0.718	0.727	0.729	0.750	0.778	0.820	0.837
LOS > Med	0.638	0.659	0.683	0.740	0.778	0.814	0.826

Table 12 F1-scores of logistic regression models developed to predict subpopulation membership of along the course of inpatient episodes.

LOS breakdown analysis on the value of training models on different LOS subpopulations. From a practical standpoint, training models in subpopulations can be beneficial and can be implemented in practice by predicting subpopulation membership with confidence (see Table 12).

Lastly, value of information has indicated that diagnoses (entered by clinicians as working diagnoses), prescriptions, medication and assessments are relevant to improve model performance. These results call out for special attention in implementing and using EHR systems, to improve downstream use of data.

5.2 Implications for hospital management and EHR systems

The key contribution of this article is a methodology for predicting different managerial variables of patients from routinely collected structured EHR data with relevance to inform hospital resource management. The performance results obtained with our methodology indicate that certain key variables relevant for hospital management can be predicted from these data. The key advantage of such systematic and holistic approach is in providing a framework approach to EHR dataset development and model-building steps that can be applied to different types of MVs and whose procedures can be partially automated, which reduces re-work of manually building models for each MV. This is particularly advantageous for technical implementation of a decision support system. In addition, models are developed from the same EHR dataset (i.e. feature definition does not have to be replicated) while allowing the flexibility to use the most appropriate feature subset (determined through feature selection) for each MV. Examples of practical utility of MV predictions in hospital management are as follows:

Data elements	Diagnoses	Procedures	DRGs	Low LOS outlier	High LOS outlier	LOS
Demographics						
Diagnoses (EHR)	▲▲		▲▲			+
Personal history						
Prescriptions		▲				
Medication	+				▲▲	▲
Assessments	+	+	+	▲▲		▲▲
Laboratory					+	

Table 13 Qualitative representation of the intensity of contribution of EHR data elements for model performance across all managerial variables (represented in columns). Results are based on F1-scores and MAE for classification and regression models, respectively. Legend: +: influence > 5%; ▲: influence > 10%; ▲▲: influence > 15%

MV1 (Diagnoses) From an admissions manager’s point of view, knowing patient diagnoses preemptively is useful in assigning patients to the correct ward and thereby allocate bed capacity and a care team which is specialized on the patient’s specific condition.

MV2 (Procedures) Predicting procedures has direct utility for efficient planning of specific procedure facilities and operating theatre time and, indirectly, improving the predictive accuracy of surgery duration. [80].

MV3 (DRGs) Knowing the DRG at early stages of care is particularly important in mitigating financial risks related to patient treatment. Also, DRGs reflect the overall resource consumption during the patient flow through the hospital, which can help develop a clinical pathway [81] for each patient’s DRG. However, there are limitations for the use of DRGs beyond a contribution margin-oriented patient scheduling because DRGs only represent a subjective function of diagnoses, procedures, and other clinical and demographic variables [82]. As a result, the combination of MVs 1, 2 and 3 may turn out very useful to improve hospital-wide patient scheduling decisions.

MV4 (Outlier episodes) Identifying outlier episodes has two implications for the bed management: if a patient is classified as an LOS outlier below the low LOS trim point, bed management may identify opportunities for freeing up bed capacity or flag early discharge with risk of readmission. On the other hand, if a patient is classified as an LOS outlier above the high LOS trim point, from a patient scheduling perspective it makes sense to accelerate, for example, the pre-surgical diagnostics phase to mitigate risks of patient LOS exceeding expected limits. However, these patient scheduling decisions should be taken carefully, always looking at the holistic picture of all hospital resources and

especially focusing in ensuring optimal patient outcomes [19].

MV5 (Length of Stay – LOS) Once a patient’s LOS is predicted, it can be linked to a patient-bed assignment model, for example, using near-real time bed modelling [83]. This is not only useful for bed capacity managers but also for patients and caregivers to plan discharge and transferred to their home or care facilities.

The associated experiments also provide insights into how to apply the methodology and seek performance optimizations. The deployment of these methods in production settings requires automatic data extraction, preprocessing and model development in order to provide relevant predictions to hospital managers in a timely manner. The inclusion of these predictions in routine workflows can help bridge communication gaps between clinical and managerial perspectives – which are often dissociated (especially in public hospitals) – and foster clinical staff awareness towards efficient operation.

Our results also draw attention to the importance of careful configuration, training and use policy to ensure EHR systems produce high-quality data that can also be reused for decision making purposes. Clinician-assigned diagnoses and assessments are particularly important in this context.

5.3 Limitations

In this study, several elements play a role in the ability to generalize these methods and results. Methodologically, there are alternatives to discretization of lab essay results, discarding features with missing values and factoring in the chronological order of episodes, which can be explored as future work. There are also other coding and DRG classification systems in use across many health systems, which could produce varying results.

The scope, quality and level of structure of source EHR data will also vary significantly across systems and will impact generalizability, however the use of EHR standards and terminologies (e.g. RxNorm [84]) can help address these limitations.

The scope of managerial variables might entail potential issues of using ICD-9-CM to describe clinical conditions (due to ambiguous/unspecified categories) and use of DRGs for direct resource management (especially being imported from countries different health systems), and may leave out other relevant measures such as prediction of materials, capacity, human resources or risk of readmission. These topics can all be subject of future work to extend the proposed methodology.

One additional potential limitation is focusing feature set construction and model development exclusively on structured EHR data. In fact, unstructured data such as medical narratives and discharge summaries continue to be highly preferred by health professionals and are pervasive in datasets. While our case study is based on a system which highly focuses in collecting health data in structured and standardized formats, this might not be the case for all hospital settings and lead to important data being left out of prediction models. The scope of structured data formats in EHR systems should be evaluated as an important factor for the success of the proposed methodology. Notwithstanding, the persisting challenges and limitations of leveraging data in unstructured formats – particularly for non-English languages – continues to motivate adoption of structured formats, which helps mitigate this potential limitation of the proposed methodology. Also, extending the proposed methodology to incorporate unstructured data in feature set construction can also be an important direction for future research.

6 Conclusions

This article presents a systematic approach to obtain predictions for the following five managerial variables:

1. Patient clinical profiles (ICD-9-CM categories),
2. Clinical procedures (ICD-9-CM codes),
3. DRG classification (AP-DRG),
4. Outlier episodes (outside of DRG LOS trim points),
and
5. Expected LOS in days.

This approach is based on information captured in structured EHR formats during the course of inpatient episodes. Prediction models are built through a standardized predictive modeling pipeline at each instant

of interest, using all EHR data available up to that instant. Furthermore, we proposed a comprehensive experimental design that addressed several research questions on the application of the proposed methodology using a dataset from a large public hospital in Portugal

This article makes three key contributions to the literature. Firstly, to the best of our knowledge, the proposed methodology is the first to employ a systematic data mining approach to predict multiple relevant managerial variables from EHR data and along the course of episodes. This includes streamlined dataset preprocessing, feature set construction and prediction model development. From a practical standpoint, this is important to enable implementation in clinical settings across multiple medical specialties, as opposed to most methodologies which are heavily tailored and prevent industrialization. Secondly, the systematic approach for temporal modeling – partitioning the full EHR dataset into filtered datasets based on timestamps instead of assuming which data are available at each episode stage – provides a new framework for obtaining predictions during the course of episodes. Thirdly, the practical case-study using a hospital dataset shows evidence that these methods can provide valuable insights to inform tactical and operational resource management decisions, specifically for different subgroups of patients/episodes which make up a significant proportion of hospital inpatient volume. These results provide tangible insights into how to leverage structured EHR data to inform proactive managerial decisions, and also provide evidence to strengthen the case for promoting collection of high-quality EHR data, including at early stages of patient stay.

Future work may include the prediction of urgency of treatment [85], linking our predictive models with prescriptive patient scheduling models [19], and predicting multiple inpatient episodes.

A Most frequent ICD 9 diagnosis codes

Order	ICD-9-CM category	Description	Order	ICD-9-CM category	Description
1	401	Essential hypertension	39	E888	Other and unspecified fall
2	276	Disorders of fluid, electrolyte, and acid-base balance	40	V46	Other dependence on machines and devices
3	250	Diabetes mellitus	41	V43	Organ or tissue replaced by other means
4	V58	Other/unspec. procedures/aftercare	42	600	Hyperplasia of prostate
5	272	Disorders of lipid metabolism	43	345	Epilepsy and recurrent seizures
6	427	Cardiac dysrhythmias	44	780	General symptoms
7	428	Heart failure	45	V60	Economic circumstances (e.g. housing)
8	486	Pneumonia, organism unspecified	46	790	Nonspecific findings on examination of blood
9	V15	Other personal history presenting hazards to health	47	595	Cystitis
10	V12	Personal history of certain other diseases	48	799	Other ill-defined and unknown causes of morbidity and mortality
11	V45	Other postprocedural states	49	593	Other disorders of kidney and ureter
12	285	Other and unspecified anemias	50	311	Depressive disorder, not elsewhere classified
13	518	Other diseases of lung	51	426	Conduction disorders
14	585	Chronic kidney disease (CKD)	52	V87	Other specified personal exposures and history presenting hazards to health
15	305	Nondependent abuse of drugs	53	290	Dementias
16	403	Hypertensive chronic kidney disease	54	287	Purpura and other hemorrhagic conditions
17	278	Overweight, obesity and other hyperalimentation	55	553	Other hernia of abdominal cavity without mention of obstruction or gangrene
18	041	Bacterial infection in conditions classified elsewhere and of unspecified site	56	404	Hypertensive heart and chronic kidney disease
19	414	Other forms of chronic ischemic heart disease	57	787	Symptoms involving digestive system
20	599	Other disorders of urethra and urinary tract	58	198	Secondary malignant neoplasm of other specified sites
21	491	Chronic bronchitis	59	412	Old myocardial infarction
22	303	Alcohol dependence syndrome	60	493	Asthma
23	584	Acute renal failure	61	535	Gastritis and duodenitis
24	V14	Personal history of allergy to medicinal agents	62	789	Other symptoms involving abdomen and pelvis
25	402	Hypertensive heart disease	63	332	Parkinsons disease
26	V49	Other conditions influencing health status	64	300	Anxiety, dissociative and somatoform disorders
27	244	Acquired hypothyroidism	65	995	Certain adverse effects not elsewhere classified
28	434	Occlusion of cerebral arteries	66	E950	Suicide and self-inflicted poisoning by solid or liquid substances
29	V10	Personal history of malignant neoplasm	67	424	Other diseases of endocardium
30	466	Acute bronchitis and bronchiolitis	68	070	Viral hepatitis
31	438	Late effects of cerebrovascular disease	69	578	Gastrointestinal hemorrhage
32	571	Chronic liver disease and cirrhosis	70	574	Cholelithiasis
33	707	Chronic ulcer of skin	71	459	Other disorders of circulatory system
34	280	Iron deficiency anemias	72	530	Diseases of esophagus
35	294	Persistent mental disorders due to conditions classified elsewhere	73	714	Rheumatoid arthritis and other inflammatory polyarthropathies
36	275	Disorders of mineral metabolism	74	112	Candidiasis
37	511	Pleurisy	75	162	Malignant neoplasm of trachea, bronchus, and lung
38	197	Secondary malignant neoplasm of respiratory and digestive systems			

Table 14 Top 75 most frequent ICD-9-CM diagnosis codes (category level), ordered by decreasing order of frequency.

B Most frequent ICD 9 procedure codes

Order	ICD-9-CM code	Description	Order	ICD-9-CM code	Description
1	90.59	Micro exam-blood NEC	39	45.13	Sm bowel endoscopy NEC
2	87.44	Routine chest x-ray	40	88.91	Mri of brain & brainstem
3	89.52	Electrocardiogram	41	54.91	Percu abdominal drainage
4	99.18	Inject/infuse electrolyt	42	44.13	Gastroscopy NEC
5	89.65	Arterial bld gas measure	43	45.23	Colonoscopy
6	99.21	Inject antibiotic	44	89.14	Electroencephalogram
7	99.29	Inject/infuse NEC	45	03.31	Spinal tap
8	93.96	Oxygen enrichment NEC	46	44.14	Closed gastric biopsy
9	91.39	Micro exam-low urin NEC	47	89.50	Ambu cardiac monitoring
10	87.03	C.A.T. scan of head	48	96.6	Enteral infus nutrit sub
11	91.33	C & s-lower urinary	49	34.91	Thoracentesis
12	90.53	C & s-blood	50	90.93	C & s-lower GI
13	88.72	Dx ultrasound-heart	51	87.49	Chest x-ray NEC
14	99.19	Inject anticoagulant	52	96.33	Gastric lavage
15	99.17	Inject insulin	53	91.19	Micro exam-periton NEC
16	88.76	Dx ultrasound-abdomen	54	38.93	Venous cath NEC
17	99.04	Packed cell transfusion	55	90.41	Bact smear-lower resp
18	93.94	Nebulizer therapy	56	91.13	C & s-peritoneum
19	87.41	C.A.T. scan of thorax	57	90.03	C & s-nervous system
20	88.01	C.A.T. scan of abdomen	58	34.04	Insert intercostal cath
21	99.23	Inject steroid	59	88.79	Dx ultrasound NEC
22	88.75	Dx ultrasound-urinary	60	33.22	Fiber-optic bronchoscopy
23	93.90	Non-invasive mech vent	61	93.99	Other resp procedures
24	90.43	C & s-lower resp	62	45.24	Flexible sigmoidoscopy
25	91.32	Culture-lower urinary	63	93.01	Functional pt evaluation
26	57.94	Insert indwelling cath	64	38.95	Ven cath renal dialysis
27	88.71	Dx ultrasound-head/neck	65	99.60	Cardiopulm resuscita NOS
28	99.28	Immunotherapy as antineo	66	51.10	Endosc retro cholangiopa
29	96.04	Insert endotracheal tube	67	90.52	Culture-blood
30	89.13	Neurologic examination	68	87.42	Thoracic tomography NEC
31	96.07	Insert gastric tube NEC	69	90.09	Micro exam-nervous NEC
32	88.77	Dx ultrasound-vascular	70	86.59	Skin closure NEC
33	39.95	Hemodialysis	71	90.49	Micro exam-lowr resp NEC
34	99.26	Inject tranquilizer	72	99.25	Inject ca chemother NEC
35	94.19	Psychia interv/eval NEC	73	99.03	Whole blood transfus NEC
36	88.19	Abdominal x-ray NEC	74	90.55	Toxicology-blood
37	96.71	Cont inv mec ven <96 hrs	75	33.24	Closed bronchial biopsy
38	88.38	Other C.A.T. scan			

Table 15 Top 75 most frequent ICD-9-CM procedure codes, ordered by decreasing order of frequency.

C Most frequent DRG codes

DRG code	Order	Description
541	1	Simple pneumonia and/or other respiratory disorders, except bronchitis or asthma, with major CC
89	2	Pneumonia and/or pleurisy with CC
127	3	Heart failure & shock
544	4	Congestive heart failure & cardiac arrhythmia with major CC
410	5	Chemotherapy
14	6	Stroke with infarct
533	7	Other nervous system disorders except transient ischemic attack, seizures, headache, with major CC
395	8	Red blood cell disorders, age >17
138	9	Cardiac arrhythmia & conduction disorder, with CC
139	10	Cardiac arrhythmia & conduction disorder, without CC
450	11	Poisoning and/or toxic effects of drugs, age >17, without CC
96	12	Bronchitis & asthma, age >17, with CC
557	13	Hepatobiliary & pancreas disorder, with major CC
320	14	Kidney and/or urinary tract infection, age >17, with CC
202	15	Cirrhosis & alcoholic hepatitis
90	16	Simple pneumonia & pleurisy, age >17, without CC
569	17	Kidney and/or urinary tract disorder, excl. renal failure, with major CC
566	18	Endocrinal, nutrition and/or metabolic disorders excl. eating disorders or cystic fibrosis, with major CC
15	19	Nonspecific cerebrovascular attack & precerebral occlusion without infarction
296	20	Nutrition & misc metabolic disorders, age >17, with CC
552	21	Digestive disorders excl. esophagitis, gastroenteritis, uncomplicated ulcer, with major CC
87	22	Pulmonary edema & respiratory failure
294	23	Diabetes, age > 35
316	24	Renal failure
88	25	Chronic obstructive pulmonary disease

Table 16 Top 25 most frequent DRG codes ordered by decreasing order of frequency.

References

1. P.R. Orszag and E.J. Emanuel. Health care reform and cost control. *New England Journal of Medicine*, 363(7):601–603, 2010.
2. Michael Carter. Diagnosis: mismanagement of resources. *OR MS TODAY*, 29(2):26–33, 2002.
3. Yasar A Ozcan. *Quantitative methods in health care management: techniques and applications*, volume 4. John Wiley & Sons, 2005.
4. Peter JH Hulshof, Nikky Kortbeek, Richard J Boucherie, Erwin W Hans, and Piet JM Bakker. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in or/ms. *Health systems*, 1(2):129–175, 2012.
5. Erwin W Hans, Mark Van Houdenhoven, and Peter JH Hulshof. A framework for healthcare planning and control. In *Handbook of healthcare system scheduling*, pages 303–320. Springer, 2012.
6. Alastair Baker. Crossing the quality chasm: a new health system for the 21st century, 2001.
7. Richard Hillestad, James Bigelow, Anthony Bower, Federico Girosi, Robin Meili, Richard Scoville, and Roger Taylor. Can electronic medical record systems transform health care? potential health benefits, savings, and costs. *Health affairs*, 24(5):1103–1117, 2005.
8. Sasikiran Kandula, Qing Zeng-Treitler, Lingji Chen, William L Salomon, and Bruce E Bray. A bootstrapping algorithm to improve cohort identification using structured data. *Journal of biomedical informatics*, 44:S63–S68, 2011.
9. Charles Safran, Meryl Bloomrosen, W Edward Hammond, Steven Labkoff, Suzanne Markel-Fox, Paul C Tang, and Don E Detmer. Toward a national framework for the secondary use of health data: an american medical informatics association white paper. *Journal of the American Medical Informatics Association*, 14(1):1–9, 2007.
10. Philip E Bourne. What big data means to me, 2014.
11. Jake Luo, Min Wu, Deepika Gopukumar, and Yiqing Zhao. Big data application in biomedical research and health care: a literature review. *Biomedical informatics insights*, 8:BII–S31559, 2016.
12. Ofir Ben-Assuli and Rema Padman. Trajectories of repeated readmissions of chronic disease patients: Risk stratification, profiling, and prediction. *MIS Quarterly*, 44(1), 2020.
13. Matthew Herland, Taghi M Khoshgoftaar, and Randall Wald. A review of data mining using big data in health informatics. *Journal of Big data*, 1(1):1–35, 2014.
14. MK Ross, Wei Wei, and L Ohno-Machado. “big data” and the electronic health record. *Yearbook of medical informatics*, 23(01):97–104, 2014.
15. Mary H Stanfill, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersh. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, 17(6):646–651, 2010.
16. Reinhard Busse, Alexander Geissler, Ain Aaviksoo, Francesc Cots, Unto Häkkinen, Conrad Kobel, Céu Mateus, Zeynep Or, Jacqueline O’Reilly, Lisbeth Serdén, et al. Diagnosis related groups in europe: moving towards transparency, efficiency, and quality in hospitals?

- Bmj*, 346:f3197, 2013.
17. D. Gartner, R. Kolisch, D.B. Neill, and R. Padman. Machine learning approaches for early drg classification and resource allocation. *INFORMS Journal on Computing*, 27(4):718–734, 2015.
 18. D. Gartner and R. Kolisch. Scheduling the hospital-wide flow of elective patients. *European Journal of Operational Research*, 233(3):689–699, 2014.
 19. D. Gartner and R. Padman. Flexible hospital-wide elective patient scheduling. *Journal of the Operational Research Society*, pages 1–15, 2019.
 20. Riccardo Bellazzi and Blaz Zupan. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, 77(2):81–97, 2008.
 21. Chun-Ling Chuang. Case-based reasoning support for liver disease diagnosis. *Artificial Intelligence in Medicine*, 53(1):15–23, 2011.
 22. Asma A Al Jarullah. Decision tree discovery for the diagnosis of type ii diabetes. In *2011 International conference on innovations in information technology*, pages 303–307. IEEE, 2011.
 23. M. Hoogendoorn, L.G. Moons, M.E. Numans, and R.J. Sips. Utilizing data mining for predictive modeling of colorectal cancer using electronic medical records. *Lecture Notes in Computer Science*, 8609:132–141, 2014.
 24. Reinier Kop, Mark Hoogendoorn, Leon MG Moons, Mattijs E Numans, and Annette ten Teije. On the advantage of using dedicated data mining techniques to predict colorectal cancer. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 133–142. Springer, 2015.
 25. Jionglin Wu, Jason Roy, and Walter F Stewart. Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, pages S106–S113, 2010.
 26. Simon Kocbek, Lawrence Cavedon, David Martinez, Christopher Bain, Chris Mac Manus, Gholamreza Haf-fari, Ingrid Zukerman, and Karin Verspoor. Text mining electronic hospital records to automatically classify admissions against disease: measuring the impact of linking data sources. *Journal of biomedical informatics*, 64:158–167, 2016.
 27. Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, 17(01):128–144, 2008.
 28. Elyne Scheurwags, Kim Luyckx, Léon Luyten, Walter Daelemans, and Tim Van den Bulcke. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. *Journal of the American Medical Informatics Association*, 23(e1):e11–e19, 2015.
 29. Maria Teresa Chiaravalloti, Roberto Guarasci, Vincenzo Lagani, Erika Pasceri, and Roberto Trunfio. A coding support system for the icd-9-cm standard. In *2014 IEEE International Conference on Healthcare Informatics*, pages 71–78. IEEE, 2014.
 30. Michael Subotin and Anthony Davis. A system for predicting icd-10-pcs codes from electronic health records. In *Proceedings of BioNLP 2014*, pages 59–67, 2014.
 31. Daniel Gartner and Rema Padman. Improving hospital-wide early resource allocation through machine learning. *Studies in health technology and informatics*, 216:315–319, 2015.
 32. Kazuya Okamoto, Toshio Uchiyama, Tadamasu Takemura, Naoto Kume, Tomohiro Kuroda, and Hiroyuki Yoshihara. Automatic selection of diagnosis procedure combination codes based on partial treatment data relative to the number of hospitalization days. *European Journal of Biomedical Informatics*, 14(1), 2018.
 33. David E Clark and Louise M Ryan. Concurrent prediction of hospital mortality and length of stay from risk factors on admission. *Health services research*, 37(3):631–645, 2002.
 34. Malcolm Faddy, Nicholas Graves, and Anthony Pettitt. Modeling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions. *Value in Health*, 12(2):309–314, 2009.
 35. David H Gustafson. Length of stay: prediction and explanation. *Health Services Research*, 3(1):12, 1968.
 36. Adrià Arboix, Joan Massons, Luís García-Eroles, Cecilia Targa, Montserrat Oliveres, and Emili Comes. Clinical predictors of prolonged hospital stay after acute stroke: relevance of medical complications. *International Journal of Clinical Medicine*, 3(06):502, 2012.
 37. Ruben L Osnabrugge, Alan M Speir, Stuart J Head, Philip G Jones, Gorav Ailawadi, Clifford E Fonner, Edwin Fonner Jr, A Pieter Kappetein, and Jeffrey B Rich. Prediction of costs and length of stay in coronary artery bypass grafting. *The Annals of thoracic surgery*, 98(4):1286–1293, 2014.
 38. Paolo Barbini, Emanuela Barbini, Simone Furini, and Gabriele Cevenini. A straightforward approach to designing a scoring system for predicting length-of-stay of cardiac surgery patients. *BMC medical informatics and decision making*, 14(1):89, 2014.
 39. Bernhard Zoller, Katharina Spanaus, Rahel Gerster, Mario Fasshauer, Paul A Stehberger, Stephanie Klinzing, Athanasios Vergopoulos, Arnold von Eckardstein, and Markus Béchir. Icg-liver test versus new biomarkers as prognostic markers for prolonged length of stay in critically ill patients—a prospective study of accuracy for prediction of length of stay in the icu. *Annals of intensive care*, 4(1):19, 2014.
 40. Asha Seth Kapadia, Wenyaw Chan, Ramesh Sachdeva, Lemuel A Moye, and Larry S Jefferson. Predicting duration of stay in a pediatric intensive care unit: A markovian approach. *European Journal of Operational Research*, 124(2):353–359, 2000.
 41. Michael Rowan, Thomas Ryan, Francis Hegarty, and Neil O’Hare. The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors. *Artificial Intelligence in Medicine*, 40(3):211–221, 2007.
 42. Nouredin Messaoudi, Jeroen De Cocker, Bernard Stockman, Leo L Bossaert, and Inez ER Rodrigus. Prediction of prolonged length of stay in the intensive care unit after cardiac surgery: The need for a multi-institutional risk scoring system. *Journal of cardiac surgery*, 24(2):127–133, 2009.
 43. PH Ong and YH Pua. A prediction model for length of stay after total and unicompartmental knee replacement. *The bone & joint journal*, 95(11):1490–1496, 2013.
 44. Evelene M Carter and Henry WW Potts. Predicting length of stay from an electronic patient record system: a primary total knee replacement example. *BMC medical informatics and decision making*, 14(1):26, 2014.
 45. Ilona WM Verburg, Nicolette F de Keizer, Evert de Jonge, and Niels Peek. Comparison of regression methods for modeling intensive care length of stay. *PloS one*, 9(10):e109684, 2014.

46. Rocco J LaFaro, Suryanarayana Pothula, Keshar Paul Kubal, Mario Emil Inchiosa, Venu M Pothula, Stanley C Yuan, David A Maerz, Lucretia Montes, Stephen M Oleszkiewicz, Albert Yusupov, et al. Neural network prediction of icu length of stay following cardiac surgery based on pre-incision variables. *PLoS One*, 10(12):e0145395, 2015.
47. Jesse Wrenn, Ian Jones, Kevin Lanaghan, Clare Bates Congdon, and Dominik Aronsky. Estimating patient's length of stay in the emergency department with an artificial neural network. In *AMIA... Annual Symposium proceedings. AMIA Symposium*, volume 2005, pages 1155–1155. American Medical Informatics Association, 2005.
48. Yang Xie, Günter Schreier, David CW Chang, Sandra Neubauer, Ying Liu, Stephen J Redmond, and Nigel H Lovell. Predicting days in hospital using health insurance claims. *IEEE journal of biomedical and health informatics*, 19(4):1224–1233, 2015.
49. Yang Xie, Günter Schreier, Michael Hoy, Ying Liu, Sandra Neubauer, David CW Chang, Stephen J Redmond, and Nigel H Lovell. Analyzing health insurance claims on different timescales to predict days in hospital. *Journal of biomedical informatics*, 60:187–196, 2016.
50. Mark Van Houdenhoven, Duy-Tien Nguyen, Marinus J Eijkemans, Ewout W Steyerberg, Hugo W Tilanus, Diederik Gommers, Gerhard Wullink, Jan Bakker, and Geert Kazemier. Optimizing intensive care capacity using individual length-of-stay prediction models. *Critical Care*, 11(2):R42, 2007.
51. Chin-Sheng Yang, Chih-Ping Wei, Chi-Chuan Yuan, and Jen-Yu Schoung. Predicting the length of hospital stay of burn patients: Comparisons of prediction accuracy among different clinical stages. *Decision Support Systems*, 50(1):325–335, 2010.
52. Zhengxing Huang, Jose M Juarez, Huilong Duan, and Haomin Li. Length of stay prediction for clinical treatment process using temporal similarity. *Expert Systems with Applications*, 40(16):6330–6339, 2013.
53. Zhengxing Huang, Wei Dong, Lei Ji, and Huilong Duan. Outcome prediction in clinical treatment processes. *Journal of medical systems*, 40(1):8, 2016.
54. Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.
55. Reinhold Haux, Christof Seggewies, Wilhelm Baldauf-Sobez, Peter Kullmann, Helmut Reichert, Laura Luedecke, and Hubert Seibold. Soarian™-workflow management applied for health care. *Methods of information in medicine*, 42(01):25–36, 2003.
56. Ramakanth Kavuluru, Anthony Rios, and Yuan Lu. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial intelligence in medicine*, 65(2):155–166, 2015.
57. Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–158. Springer, 2011.
58. Steven J Littig and Mark W Isken. Short term hospital occupancy prediction. *Health care management science*, 10(1):47–66, 2007.
59. Ministério da saúde, portaria n.o 163/2013, (2013) 2495–2606.
60. Kenney Ng, Amol Ghoting, Steven R Steinhubl, Walter F Stewart, Bradley Malin, and Jimeng Sun. Paramo: a parallel predictive modeling platform for healthcare analytic research using electronic health records. *Journal of biomedical informatics*, 48:160–170, 2014.
61. Shanshan Qiu, Ratna Babu Chinnam, Alper Murat, Basam Batarse, Hakimuddin Neemuchwala, and Will Jordan. A cost sensitive inpatient bed reservation approach to reduce emergency department boarding times. *Health care management science*, 18(1):67–85, 2015.
62. Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.
63. Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.
64. Huan Liu, Hiroshi Motoda, Rudy Setiono, and Zheng Zhao. Feature selection: An ever evolving frontier in data mining. In *Feature Selection in Data Mining*, pages 4–13, 2010.
65. Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research*, 5(Oct):1205–1224, 2004.
66. Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (8):1226–1238, 2005.
67. Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *Icml*, volume 97, page 35, 1997.
68. Norman R Draper and Harry Smith. *Applied regression analysis*, volume 326. John Wiley & Sons, 1998.
69. Leo Breiman. *Classification and regression trees*. Routledge, 2017.
70. Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, 45(1):S199–S209, 2009.
71. David W. Hosmer and Stanley Lemeshow. *Applied logistic regression*. Wiley New York, 2000.
72. Riyaz Sikora et al. A modified stacking ensemble machine learning algorithm using genetic algorithms. In *Handbook of Research on Organizational Transformations through Big Data Analytics*, pages 43–53. IGI Global, 2015.
73. Khaled Fawagreh, Mohamed Medhat Gaber, and Eyad Elyan. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal*, 2(1):602–609, 2014.
74. Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.
75. Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of machine learning research*, 13(Jan):27–66, 2012.
76. Serguei VS Pakhomov, James D Buntrock, and Christopher G Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association*, 13(5):516–525, 2006.
77. Richárd Farkas and György Szarvas. Automatic construction of rule-based icd-9-cm coding systems. In *BMC bioinformatics*, volume 9, page S10. BioMed Central, 2008.
78. Jian-Wu Xu, Shipeng Yu, Jinbo Bi, Lucian Vlad Lita, Radu Stefan Niculescu, and R Bharat Rao. Automatic medical coding of patient records via weighted ridge regression. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pages 260–

265. IEEE, 2007.
79. Yan Yan, Glenn Fung, Jennifer G Dy, and Romer Rosales. Medical coding classification by leveraging inter-code relationships. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 193–202. ACM, 2010.
 80. Marinus JC Eijkemans, Mark Van Houdenhoven, Tien Nguyen, Eric Boersma, Ewout W Steyerberg, and Geert Kazemier. Predicting the unpredictable: a new prediction model for operating room times using individual characteristics and the surgeon’s estimate. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 112(1):41–49, 2010.
 81. E. Aspland, D. Gartner, and P. Harper. Clinical pathway modelling: a literature review. *Health Systems*, pages 1–23, 2019.
 82. D. Gartner. Scheduling the hospital-wide flow of elective patients. *Springer Lecture Notes in Economics and Mathematical Systems*, 2015. Heidelberg.
 83. T. England, D. Gartner, E. Ostler, P. Harper, D. Behrens, J. Boulton, D. Bull, C. Cordeaux, I. Jenkins, and F. Lindsay. Near real-time bed modelling feasibility study. *Journal of Simulation*, pages 1–12, 2019.
 84. Simon Liu, Wei Ma, Robin Moore, Vikraman Ganesan, and Stuart Nelson. Rxnorm: prescription for electronic drug information exchange. *IT professional*, 7(5):17–23, 2005.
 85. Jonas Krämer, Jonas Schreyögg, and Reinhard Busse. Classification of hospital admissions into emergency and elective care: a machine learning approach. *Health care management science*, 22(1):85–105, 2019.