Information & Management xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Information & Management

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

Data analytics for the sustainable use of resources in hospitals: Predicting the length of stay for patients with chronic diseases

Hamed M. Zolbanin[a,*], Behrooz Davazdahemami[b], Dursun Delen[c], Amir Hassan Zadeh[d]

[a] Department of MIS, Operations Management, and Decision Sciences, University of Dayton, Dayton, OH, USA
[b] Department of Information Technology and Supply Chain Management, College of Business and Economics, University of Wisconsin Whitewater, Whitewater, WI, 53190, USA
[c] Department of Management Science and Information Systems, Center for Health Systems Innovation, Spears School of Business, Oklahoma State University, Tulsa, OK, 74106, USA
[d] Department of Information Systems and Supply Chain Management, Raj Soin College of Business, Wright State University, Dayton, OH, 45435, USA

A R T I C L E I N F O

Keywords:
Data analytics
Length of hospital stay
Deep learning
Temporal evaluation
Sustainability

A B S T R A C T

Various factors are behind the forces that drive hospitals toward more sustainable operations. Hospitals contracting with Medicare, for instance, are reimbursed for the procedures performed, regardless of the number of days that patients stay in the hospital. This reimbursement structure has incentivized hospitals to use their resources (such as their beds) more efficiently to maximize revenues. One way hospitals can improve bed utilization is by predicting patients' length of stay (LOS) at the time of admission, the benefits of which extend to employees, communities, and the patients themselves. In this paper, we employ a data analytics approach to develop and test a deep learning neural network to predict LOS for patients with chronic obstructive pulmonary disease (COPD) and pneumonia. The theoretical contribution of our effort is that it identifies variables related to patients' prior admissions as important factors in the prediction of LOS in hospitals, thereby revising the current paradigm in which patients' medical histories are rarely considered for the prediction of LOS. The methodological contributions of our work include the development of a data engineering methodology to augment the data sets, prediction of LOS as a numerical (rather than a binary) variable, temporal evaluation of the training and validation data sets, and a significant improvement in the accuracy of predicting LOS for COPD and pneumonia inpatients. Our evaluations show that variables related to patients' previous admissions are the main driver of the deep network's superior performance in predicting the LOS as a numerical variable. Using the assessment criteria introduced in prior studies (i.e., ± 2 days and ± 3 days tolerance), our models are able to predict the length of hospital stay with 86 % and 91 % accuracy for the COPD data set, and with 74 % and 85 % accuracy for the pneumonia data set. Hence, our effort could help hospitals serve a larger number of patients with a fixed amount of resources, thereby reducing their environmental footprint while increasing their revenue, as well as their patients' satisfaction.

1. Introduction

In line with legislative changes to hospital Medicare reimbursement, referred to as the inpatient prospective payment system (IPPS), hospitals that have contracted with Medicare to provide acute inpatient care are paid a predetermined rate as payment in full. Hospitals receive standardized payments via IPPS for the procedures performed, regardless of the number of days that patients stay in the hospital. Thus, this reimbursement structure has incentivized hospitals to use their resources (e.g., beds) more efficiently in order to maximize Medicare revenue [1,2]. To improve bed utilization, hospitals must be able to predict, with acceptable accuracy, a patient's length of stay (LOS) at the time of admission.

The benefits of predicting LOS in hospitals are not solely economic, but also extend to other aspects of care, and of course, to the patients themselves. For instance, predicting LOS enables hospitals to identify patients who are at a higher risk for an extended stay, which can then be used to optimize their treatment plans [3,4] or can enable early interventions in order to prevent hospital-acquired infections and other complications [5]. Similarly, predicting LOS allows for planned early discharge to the patient's home or less expensive health care facilities that are supported by community nurses and doctors [6–8]. The

* Corresponding author at: University of Dayton, Department of MIS, OM, and DS, 300 College Park, Dayton, OH, 45469-2130, USA.
E-mail address: hmzolbanin@udayton.edu (H.M. Zolbanin).

https://doi.org/10.1016/j.im.2020.103282
Received 18 February 2019; Received in revised form 9 February 2020; Accepted 11 February 2020

broader impact of predicting hospital LOS, however, is on communities, through the better management of resources and the reduction of hospital waste [9–13]. Specifically, accurately forecasting patients' LOS as a key indicator of inpatient resource consumption [14–17] gives hospitals the means to accommodate more patients with the same volume of resources [1,3], thereby decreasing their ecological footprint. Therefore, predicting patients' LOS equips health care delivery organizations with a multitude of advantages that stretch from improved patient satisfaction [6,12,18] to more efficient utilization of manpower and facilities [5], reduced treatment costs (Ying [19]), and improved sustainability [11]. In recent years, a major contributor to realizing these improvements in health care outcomes has been the increasing use of [big] data analytics [20,21].

Regarding the capabilities of analytics in health care ([22,23]) and the economic and environmental impacts of extended LOS in hospitals, this paper employs a data analytic approach to develop a model for the early prediction of LOS for patients with chronic diseases. More specifically, we illustrate how a focus on preprocessing large electronic medical records (EMR), rather than on building more sophisticated models or including more variables, can significantly improve the performance of predictive models for hospital LOS. We complement our focus on data preparation and data engineering by employing an advanced deep learning model that is capable of extracting the elusive features of data sets to further improve predictions.

This study, therefore, has four main contributions (where the first three are methodological, and the last one is theoretical). First, unlike most (if not all) prior studies, in which the temporal order of events in hospital encounters is disregarded and a random split is used to create the training and validation data sets, this effort uses temporal evaluation for predicting hospital LOS. Second, it improves the accuracy of predicting LOS as a continuous variable (i.e., predicting the exact length of hospital stay, as opposed to using a threshold for predicting short vs. long stays) by proposing a data engineering methodology that creates new variables from the existing ones. Third, it illustrates how the use of cutting-edge analytics techniques (i.e., deep learning) can result in the development of more sustainable operations in hospitals by enabling more accurate predictions. Fourth, it identifies variables related to patients' previous admissions as important factors in predicting LOS in chronic diseases. This is an important contribution because it underscores the significance of medical history in predicting LOS; as a result, this study revises the current paradigm that uses variables related only to current admissions for the prediction of LOS.

The paper proceeds as follows: In the next section, we review the literature on applying data analytics to predict LOS. In the subsequent section, we describe the data sets that we use in our study. Next, we explain our data engineering methodology, followed by a description of the deep learning model we design and implement to predict LOS. Subsequently, we provide an analytical evaluation of the predictive model and discuss the paper's contributions and implications for practice and future research.

2. Prior work

Increasing data analytics applications in health care have the potential to revolutionize hospitals by enabling them to serve more patients with the same amount of resources, thereby reducing the impact of health care on the environment [24]. The benefits of sustainable health care, which are mainly realized via quality and financial improvements, are not limited to the environment [11]. Indeed, sustainable health care may also transpire in other forms [11,25,26]: customer-oriented sustainability (such as an enhanced quality of patient care, increased patient satisfaction, reduced medical bills); employee-oriented sustainability (such as improvement in professionals' job satisfaction); or community-oriented sustainability (such as saving energy and materials and reducing pollution). By making an impact on all of these dimensions, predicting hospital LOS is one way that the health care industry can enhance patient outcomes and move toward more sustainable operations.

Prior studies on predicting LOS – with a predictive rather than a descriptive setup – have mainly used one of two approaches: nominal (mostly binary) or continuous prediction of the target variable (i.e., LOS). A majority of these studies have expressed LOS as a binary variable, with various operationalizations of how the two levels of the variable are assigned. In one study, for instance, the problem was operationalized as "early" versus "end of the day" discharge of patients [6]. In others, prolonged LOS was predicted, with different values - depending on the index condition - as the cutoff point for the target variable ([4,9,18,27–30] [19];). Another group of studies used a descriptive approach to identify factors that were highly associated with LOS [15,31–37].

In contrast, fewer studies have tried to predict LOS as a numerical variable, most probably because doing so for variables with a small range of values is extremely elusive. In a recent study, a regression tree model was used to predict LOS for patients with congestive heart failure [38]. In other studies, artificial neural networks were used to predict LOS for three heart diseases [12]; regression models were employed to predict this variable in total knee replacement [39]; and random forest models were used to predict LOS for patients in a large hospital group [3] and for hip-fracture patients [40]. However, we did not come across any studies that used deep learning, whose state-of-the-art evolution has shown great promise for discovering intricate structures in high-dimensional data [41].

In addition to the operationalization of LOS and the techniques used to predict it, the current literature has a gap in incorporating patients' medical histories into the prediction of LOS. While medical histories are captured in EMR, they are usually stored at the transaction level. Therefore, due to various obstacles, such as heterogeneity among different health systems [42], locating patients' previous hospital visits and aggregating them at the patient level require extensive data preprocessing and/or engineering [43]. As a result, the impact of patients' medical histories on LOS in chronic diseases (and the extent of this impact) remains an open question.

Hence, our overall evaluation of the prior literature, which is summarized in Appendix A, reveals a few possibilities for improvement. First, predicting LOS as a binary or nominal variable provides limited utility to health care delivery organizations because it can only specify whether a patient will stay in the hospital longer or shorter than a predetermined cutoff. Second, several binary models, such as Gholipour et al. [28] and Launay et al. [29], usually have a much higher specificity than sensitivity. In other words, they perform fairly well in predicting cases in which the patient stays in the hospital for shorter periods, but do not have sufficient discriminative power for patients whose LOS is longer than the cutoff. Third, some studies, such as Hachesu et al. [18], report the results for the training data sets, which undermines their generalizability to other data. Fourth, almost all studies (including those that consider LOS as a numerical variable) use a random split to assign data to the training and testing sets. This is problematic because using a random split would mean that the data on patients' future hospitalizations are used to predict their LOS for earlier hospital stays, which undermines the applicability of such models. Fifth, most (if not all) studies have not considered patients' medical histories, especially chronic diseases, in the prediction of hospital LOS. In particular, the contribution of adding information from previous hospitalizations (both at the patient and population level) to the prediction of LOS is unknown.

To address these issues, we use temporal evaluation and a numerical operationalization of LOS to predict this variable for patients with chronic diseases. More specifically, we use earlier hospital admissions to train a deep neural network and use latter visits to evaluate the model's performance. As we discuss later in the paper, we find that a patient's average LOS prior to the current hospitalization, as well as a few other variables related to previous hospital admissions, contribute the most to the accuracy of our predictive model. Therefore, our study contributes to the health care analytics literature by illustrating the importance of patients' medical histories in the early prediction of hospital LOS. Likewise, via a more accurate prediction of LOS at the time of admission, it advances sustainable health care practices by reducing the consumption of resources, and thus, the amount of waste generated by hospitals.

To demonstrate these contributions, we employ a data engineering methodology to aggregate prior visits at the patient and data set levels, and we develop new variables based on these aggregations. Before explaining our methodology, we describe the data sets used for this study in the following section.

3. The data

The challenges and computational constraints associated with the storage, processing, and aggregation of electronic health records minimize volume as the exclusive indicator of big data in health care [44]. Instead, the complex methods, resource-intensive computations, and time-consuming adjustments that are required to assure the veracity of health data – which is critical in enabling value generation – are the main data challenges in this domain [44]. As a result, the computations needed to preprocess medical records with several thousand encounters may well qualify as applications of big data analytics in health care [43].

The data sets we use in this study are derived from a large data warehouse composed of more than 2.5 Terabytes of clinical data from computerized physician order entry (CPOE) systems (i.e., physician's notes and prescriptions, medical imaging, laboratory, pharmacy, insurance, and administrative data) and other patient data in an electronic format. The data warehouse includes nearly 380 million unique hospital encounters, generated from visits by more than 63 million unique patients over a 15-year period between 2000 and 2015. We created two data sets by limiting the data to inpatient records whose primary diagnosis was pneumonia or chronic obstructive pulmonary disease (COPD) and allied conditions. Approximately 23 % of patients in the COPD data set and more than 50 % of patients in the pneumonia data set suffered from two or more conditions at the time of admission. These diseases are among the chronic conditions that have been under heightened scrutiny by the Centers for Medicare and Medicaid Services (CMS) within the past several years through such initiatives as the hospital readmission reduction program (HRRP).

Additionally, we filtered each of the data files to include more recent encounters for two reasons. First, the sample statistics obtained from the more recent encounters are more compatible with the current population parameters in the United States. Second, due to the computing power of our devices, we needed to limit the size of the data sets so that our preprocessing procedures would run in a reasonable amount of time. Consequently, for the COPD data, we retained encounters whose date/time of admission and discharge were between January 1 and December 31 of 2015, respectively. For the pneumonia data set, the timespan of the data was between January 1, 2010 and December 31, 2015. In the last data preparation step, using a cutoff of four standard deviations above the mean value of LOS, we identified and removed the outliers from each of the data sets. As a result, 1769 and 841 records were deleted from the COPD and pneumonia data sets, representing 2% and 1.31 % of the records in the respective files.

The final COPD data set contains 86,338 encounters from 73,901 unique patients admitted to 182 hospitals in various geographical regions of the United States. The average number of hospital stays is 1.17, and the average LOS is 5.15 days. The pneumonia data set contains 63,185 encounters from 53,476 unique patients admitted to 202 hospitals. The average number of admissions is 1.18, and the average LOS is 8.31 days (the distribution of LOS in each of the data sets is given in Appendix B). Table 1 summarizes the demographics of the data sets. A description of the variables and their frequencies is given in Table 2.

Table 1
Demographics of the data.

Variable		COPD	Pneumonia
Gender	Female	59.68 %	49.8 %
	Male	40.32 %	50.2 %
Age	Mean	56.52	61.23
	Std Dev	22.15	23.43
Age (Female)	Mean	56.67	62.13
	Std Dev	21.24	23.25
Age (Male)	Mean	56.31	60.34
	Std Dev	23.43	23.57
Race	White	70.35 %	72.16 %
	Black	20.10 %	20.17 %
	Other	9.55 %	7.67 %
Marital Status	Divorced	12.15 %	9.97 %
	Married	34.48 %	34.11 %
	Single	34.61 %	30.70 %
	Widowed	14.57 %	19.96 %
	Other	4.19 %	5.26 %
Hospital's Census Region	Midwest	26.77 %	22.92 %
	Northeast	19.68 %	30.93 %
	South	32.24 %	38.15 %
	West	21.31 %	8.01 %

Since the purpose of our study is to predict the length of a hospital stay when a patient is admitted, we only consider variables whose values are known at the time of admission. Consequently, we exclude all variables that are populated at discharge (e.g., total charges and discharge disposition). Additionally, due to scant variability in some of the variables (e.g., acute vs. non-acute status) or their excessive missingness (e.g., patient's weight), we do not include them in our analyses.

After describing the data set and its variables, we now turn our attention to the methodology we use to develop our predictive model.

4. Methodology

Besides a chronic disease's long-lasting nature, which in and of itself increases the extent to which hospital resources are used, the severity of these persistent conditions is an additional factor contributing to the consumption of resources in hospitals [45,46,37,47]. Consequently, incorporating information on the history of hospital use can lead to more accurate predictive models for patients with chronic diseases [43]. In this section, we explain our methodology by dividing the contents into two parts. The first part explains our data engineering approach, in which we extract a piece of the patients' historical information from the electronic medical records and integrate them with the transactional data. As we will discuss later, this integration significantly improves the performance of the predictive models. The second part describes our model-building endeavor and presents the settings we used to develop our predictive tool. Fig. 1 provides a depiction of our methodology.

4.1. Data engineering

As we mentioned previously, having a more comprehensive picture of a patient's health can lead to the development of more accurate predictive models in chronic diseases. One way to obtain that information, in part, is through extracting information from patients' previous visits and aggregating that information with their current hospital stays. In this way, we are able to create a connection between patients' various visits rather than treating them as independent transactions. Regarding the dependent variable in this study (i.e., LOS), we create new variables in each encounter that represent a patient's

Table 2
Variable Definitions and Frequencies.

Variable	Description	Type	Levels	Data	
				COPD	Pneumonia
Encounter					
Admission source	How the patient was referred to the hospital	Nominal	Clinic referral	8.22 %	3.70 %
			Emergency room	16.15 %	19.90 %
			Physician referral	45.39 %	51.72 %
			Transfer from another health care facility	8.10 %	2.16 %
			Other	22.14 %	22.52 %
Admission type	Medical emergency of the admission	Nominal	Elective	18.40 %	5.73 %
			Emergency	59.80 %	66.96 %
			Urgent	11.14 %	9.97 %
			Other	10.66 %	17.34 %
Payer	How the patient charges will be paid	Nominal	Medicaid	15.58 %	13.04 %
			Medicare	42.45 %	51.11 %
			Blue Cross – Blue Shield	4.99 %	3.78 %
			Other commercial payer	12.15 %	6.04 %
			Self-pay	2.68 %	2.28 %
			Other	22.15 %	23.75 %
Is readmission	Whether the current encounter is a 30-day readmission	Binary	No	80.36 %	68.35 %
			Yes	19.64 %	31.65 %
First admission	Whether the current encounter is the patient's first admission in the data set	Binary	No	43.86 %	16.48 %
			Yes	56.14 %	83.52 %
Length of Stay	The number of days a patient stays in the hospital	Numerical	Mean (Std. Dev.)	5.15 (5.22)	8.31 (11.33)
Diagnosis					
Diagnosis code	Diagnosis for the encounter (up to 3 diagnoses)	Nominal	ICD-9 Code for a condition related to heart failure, pneumonia, or COPD	NA	NA
Hospital					
Bed size range	Size of the hospital	Nominal	1-5	4.41%	1.45 %
			6-99	7.30%	7.86 %
			100–199	12.34%	12.82 %
			200-299	19.25%	20.33 %
			300-499	27.68%	30.18 %
			500+	29.02 %	27.35 %
Teaching facility	Whether the hospital is a teaching facility	Binary	Yes	74.34 %	73.26 %
			No	24.68 %	26.52 %
			Missing	0.98 %	0.22 %
Cath lab full indicator	Whether the hospital has a full Catheterization laboratory	Binary	Yes	82.18 %	84.91 %
			No	14.70 %	12.52 %
			Missing	3.12 %	2.57 %

average LOS prior to their current hospitalization, as well as the average LOS of all visits in the data set whose time of discharge was earlier than the current encounter's time of admission. We perform the data engineering part of this study in two steps: extraction of historical data and approximation of missing information.

4.1.1. Extraction of historical data

Since previous research has found that patients with a higher severity of illness are more likely to stay longer in the hospital and consume greater health care resources [48], the inclusion of a history of the patients' health status and health care use could improve the performance of predictive analytics models. In this section, we use a simple mathematical notation to explain the process through which patients' historical data are extracted. Let P_i be patient *i*, $P_i[V_j]$ represent the j^{th} hospital visit of P_i, and $P_i[V_j[X]]$ represent the value of variable X for the j^{th} visit of P_i. Then, we show the admission and discharge date/time of $P_i[V_j]$ with $P_i[V_j[T_{admission}]]$ and $P_i[V_j[T_{discharge}]]$, respectively. The two calculated variables that we incorporate into our data sets represent each patient's average individual LOS over their previous admissions, as well as the overall average LOS across all admissions (by any patient) that occurred before the current admission of the current patient. The former of these variables provides an overall picture of a patient's health status, and the latter allows for a comparison between the health status of any patient with that of the average patient with a similar illness. The values of these new variables are determined according to Eq. 1 and Eq. 2.[1]

$$P_i[V_j[avg\_patient\_los]] = \frac{\sum_{k=1}^{j} P_i[V_k[LOS]]}{j} \tag{1}$$

$$P_n[V_m[avg\_total\_los]] = \frac{\sum P_i[V_j[LOS]]}{count(P_i[V_j])} \; where \; P_i[V_j[T_{discharge}]] < P_n[V_m[T_{admission}]] \tag{2}$$

Because the admission and discharge times of all visits are known, *avg_patient_los* will be assigned non-missing values for all repeated admissions of any patient. However, since no personal records exist before a patient's first hospitalization, this variable will have a missing value in the first admission of each patient. Therefore, we need to address the missingness problem in the first calculated variable. We explain our approach to handling this issue in the following subsection. Missingness, however, is of no concern for *avg_total_los,* since only the first record in the data - in chronological order of admission - does not have any preceding records.

[1] We use the same logic to create two additional variables representing each patient's average time between encounters (*pt_avg_readmission*) and the overall average time between encounters (*avg_readmission*). In the interest of brevity, we explain the methodology for one set of variables only.

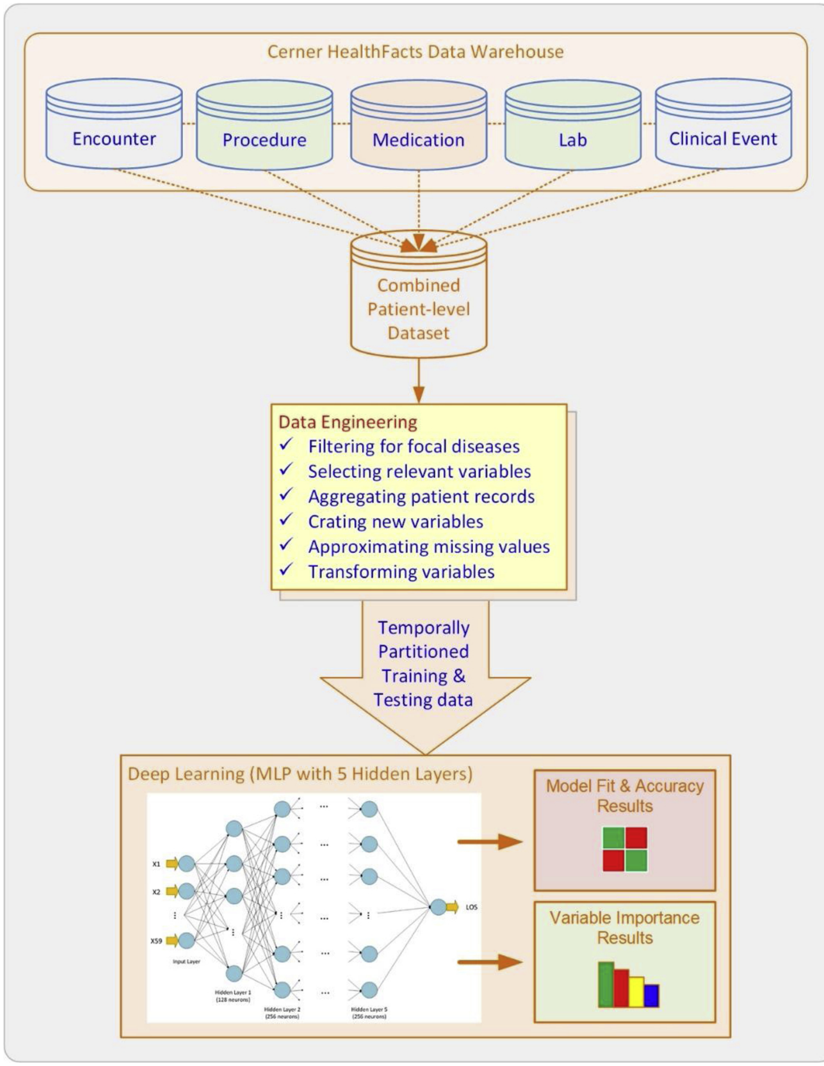


Fig. 1. Research methodology depicted as a workflow.

4.1.2. Approximation of missing information

Although we are not performing a survival analysis in this study, the data set we use may, in a way, look similar to survival data: some of the observations representing previous (or even future) hospital encounters of any specific patient may have been censored. In fact, this characteristic is common in most data sets extracted from electronic medical records. However, in contrast to survival data, in which the response variable is the waiting time until the occurrence of a well-defined event [49], the response variable in our study is not dichotomous. In addition, censoring in our study has a different meaning than a non-occurrence of the event of interest at the time of analysis. Therefore, it is appropriate to deal with this type of data incompleteness differently from how it is commonly dealt with in survival analysis.

We use the term *censoring* to refer to the situation in which the time of some[2] patients' previous or prospective hospital visits falls outside the beginning or the end of the electronic medical records we extracted from the data warehouse. Based on this definition, if a patient was admitted to the hospital some time before the beginning of the data set or was discharged after the end of the data set, the records representing those encounters are censored from our data sets. As a result, our data sets contain only those hospital stays in which both the admission and discharge times fall within the data sets' beginning and end. Some possible scenarios are illustrated in Fig. 2.

Because we are interested in patients' historical data, we only need to deal with the loss of information due to left censoring. In other words, right-censored records contain information that is beyond our study's time limits, and therefore, are not estimated in our data engineering effort. Left censoring, on the other hand, results in two types of missingness, or information gaps, that need to be handled. Type I occurs when a patient's record is dropped because its time of admission is before the study's onset, whereas Type II pertains to the situation in which a patient does not have any admission records until some time in the middle of the data set. The latter scenario itself can happen under various circumstances. For instance, a hospital record may indeed represent a patient's first ever hospital stay for that index disease, or the patient's prior admissions may have occurred in hospitals that did not contribute to the data collected in the data warehouse. While the loss of information due to censoring can happen in different forms and for various reasons, the data incompleteness resulting from it creates a common problem for data analysis. Regarding the focus of this study, which is on developing a predictive model, we do not differentiate between the various causes of missingness; hence, we employ a data analytic strategy to cope with this problem.

To replace the missing values in the calculated variables that extract and aggregate the historical information of all patients in the data set, we use a decision tree setting to estimate a value for each variable with missing data by analyzing those variables as the target. Consequently,

[2] In fact, most (if not all) patients' data will be either censored from the left (beginning) or right (end). For many, the data may be censored on both sides.

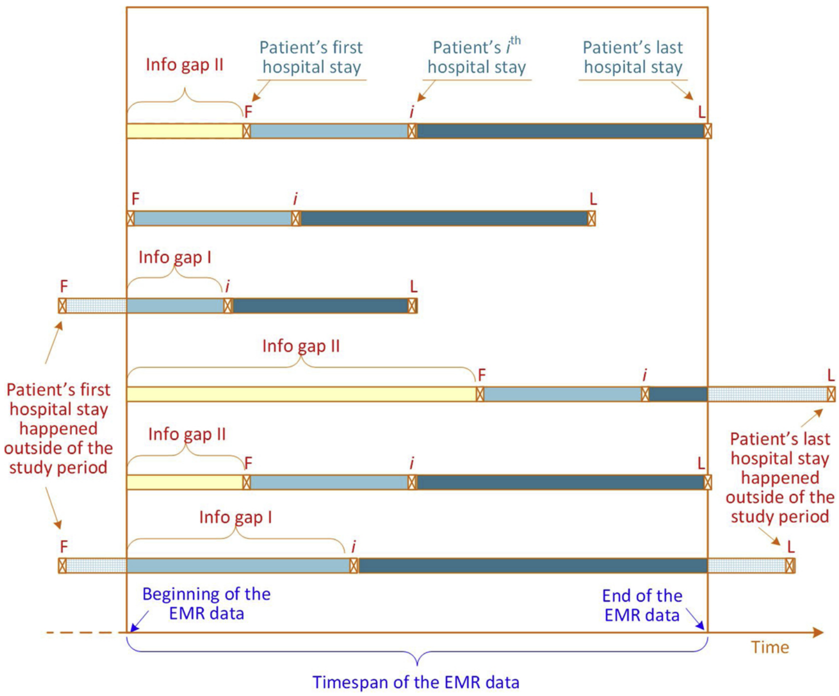


Fig. 2. Timing of patients' hospital stays versus timespan of EMR data.

Table 3
Parameter settings of the decision tree.

Parameter	Description	Value
Leaf Size	Specifies the smallest number of training observations that a leaf can have.	5
Maximum Branch	Restricts the number of subsets that a splitting rule can produce to the specified number or fewer. For example, a value of 2 results in binary trees.	2
Maximum Depth	Specifies the maximum number of node generations. The original node, generation 0, is called the root node. The children of the root node are the first generation.	6
Number of Rules	Specifies how many rules are saved with each node. The tree uses only one rule; the remaining rules are saved for comparison.	5
Number of Surrogate Rules	Specifies the maximum number of surrogate rules that are sought in each non-leaf node. A surrogate rule is a backup to the main splitting rule. When the main splitting rule relies on an input whose value is missing, the first surrogate rule is invoked.	2

in this setting, all other independent variables are used as predictors to provide an estimate for the missing values in the calculated variable. As an example, suppose a data set has ten independent variables, x_1 to x_{10}, and the goal is to impute the missing values in x_1 using the other predictors. The decision tree method would then use x_1 as the dependent variable and x_2-x_{10} as the independent variables. These nine predictors are ranked by their ability to predict x_1, and those that do no better than the marginal distribution of x_1 are excluded from further consideration. The variable with the best prediction performance for x_1 is used to replace the missing values for x_1. If the best predictor of x_1 has missing values, the best surrogate variable with non-missing data is used in its place [50]. Since this approach approximates missing values by using information from other input variables, it provides far better results than simply using a fixed value, such as the variable mean or median [51]. The parameter settings of the decision tree used to replace the missing values in the calculated variables are shown in Table 3. These values were selected after experimenting with several different options for each parameter.

Before explaining our model development methodology, in the next subsection, we provide an analysis of the computational complexity of the proposed data engineering approach and explain why it qualifies as a big data analytics application.

4.1.3. Computational complexity

We determine the computational complexity of the proposed data engineering approach using the mathematical big O notation. Let f and g be two functions defined on some subset of real numbers. Then, we say $f(x) = O(g(x))$ as $x \to \infty$ if, and only if, there is a positive constant M such that for all sufficiently large values of x, the absolute value of $f(x)$ is smaller than or equal to the absolute value of $g(x)$. That is, $f(x) = O(g(x))$ if, and only if, there exists a real number M and a real number x_0 such that $|f(x)| \le M\,|g(x)|$ for all $x \ge x_0$. Clearly, if $f(x) = O(g(x))$ and $g(x) = O(h(x))$, then $f(x) = O(h(x))$.

Now, let n be the total number of records, c be the number of variables in the data set, m show the number of unique patients, and $\bar{v}$ represent the average number of hospital visits in the electronic medical data. Then, because some patients have multiple admissions, we can conclude that $m < n$ and $m\bar{v} \approx n$. Moreover, assume that Fig. 3 illustrates a snapshot of the data sets we used in this study. Then, obtaining the average LOS for each patient would be $O(m\bar{v}^2) = O(n\bar{v})$, extracting the average total LOS would be O(n^2), and estimating the average LOS for each patient's first encounter would be O(cm), as we explain next.

As can be seen in Fig. 3, except for the first hospital stay of each patient, calculating the average LOS before a patient's i^{th} visit requires accessing all of her prior hospitalizations, which are equal to i -1 records. Therefore, for a patient with $\bar{v}$ records, the total number of operations will be $\sum_{j=1}^{\bar{v}-1} j = \frac{\bar{v}(\bar{v}-1)}{2}$. This process is repeated for all patients; therefore, $m\,\frac{\bar{v}(\bar{v}-1)}{2}$ records, on average, will be retrieved to create the individual-level computed variable. Consequently, the creation of *avg_patient_los* results in operations at the order of $O(m\,\bar{v}^{\,2})$, which in turn, equals $O(n\,\bar{v}\,)$.

Record Number	Patient Number	Variable 1	Variable 2	...	Variable k	Visit Number
1	Patient 1	...	...	...	...	1
2	Patient 1	...	...	...	...	2
...	...	...	...	...	...	...
$\overline{v}$	Patient 1	...	...	...	...	$\overline{v}$
$\overline{v}$ + 1	Patient 2	...	...	...	...	1
$\overline{v}$ + 2	Patient 2	...	...	...	...	2
...	...	...	...	...	...	...
2$\overline{v}$	Patient 2	...	...	...	...	$\overline{v}$
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
n - ($\overline{v}$ - 1)	Patient m	...	...	...	...	1
n - ($\overline{v}$ - 2)	Patient m	...	...	...	...	2
...	...	...	...	...	...	...
n - 1	...	...	...	...	...	$\overline{v}$ -1
n	Patient m	...	...	...	...	$\overline{v}$

Fig. 3. A snapshot of the data set.

To obtain the computation complexity of Eq. 2 (used to create *avg_total_los),* we need to pay attention to its difference with the computation performed in Eq.1. In the first equation, we are only interested in prior admissions of a certain patient, whereas in Eq.2, we want to calculate the average LOS across *all* visits by *all* patients prior to the current record. Thus, for each record in the data, we need to retrieve a fraction of the entire data set, and as the date/time of the admission approaches the end of the data, this fraction becomes increasingly larger. As a result, if we sort the data set by the increasing order of the discharge date, the procedure represented in Eq. 2 retrieves $^1/_n$ of the data for the first record, $^2/_n$ of the data for the second record, and $^n/_n$ or all of the data for the last record to calculate the total average LOS across all admissions that occurred prior to a given record. In other words, for the i^{th} record in the sorted data set, Eq. 2 retrieves *i* records to compute *avg_total_los*. Hence, the time complexity of Eq. 2 is $\sum_{i=1}^{n} i = \frac{n(n+1)}{2} = O(n^2)$.

Finally, to estimate the average LOS in the first admission of each patient - where Eq.1 returns a missing value due to left censoring - the decision tree will, at most, peek at *(c-1)* variables to find the best replacement. Therefore, the procedure used to approximate the missing information is *O(cm).*[3]

It follows from our analyses that the overall complexity of the data preparation steps we employed above is $O(n^2)$, where *n* is the number of records in the data. However, as we will discuss later in the paper, *avg_patient_los* is the main driver regarding the performance of the predictive models we build in this effort. As a result, it is possible to create a balance between the computational complexity of the data engineering steps and the models' predictive performance by forgoing the creation of the *avg_total_los* variable. In that case, the computational complexity of the methodology we use in this study would be reduced to $O(n\overline{v})$.

In either case, the complexities of working with health data, including the number of records that need to be accessed or processed, the number of diseases and procedures, and the time-consuming adjustments that are needed to ensure data quality [44], require big data technologies to obtain acceptable results in a reasonable amount of time. Additionally, the size of the databases or data warehouses that host electronic medical records increases significantly every few months, creating the need to rerun several procedures of at least *O(n)* on a regular basis for data sets that encompass several million records. It is obvious that such operations will eventually exhaust average computers and will require big data or distributed processing in the end.

4.2. Model building

As pointed out earlier, we use a deep learning neural network – a representation learning method [52] - to build our predictive model. Representation learning techniques are a type of machine learning in which the emphasis is on learning and discovering the data features, in addition to finding the mapping from those features to the output. Fig. 4 highlights the differences in the steps of a typical deep learning model with those of classic machine learning algorithms. As shown in this flowchart, deep learning enables the computer to derive, from simple concepts, some complex features whose discovery can be very laborious for humans. It then maps those advanced features to the output.

From a methodological viewpoint, although deep learning is generally deemed to be a new area in machine learning, it is in fact an extension to the idea of neural networks that can deal with more sophisticated tasks, and can handle larger data sets with many variables at the expense of greater computational effort.

Multilayer perceptron (MLP) networks, also known as deep feedforward networks, are the most general type of deep networks. These networks are large-scale neural networks that can contain many layers of neurons and handle tensors as input. The types and characteristics of the network elements (i.e., weight functions, transfer functions, etc.) are the same as in the standard neural network models. These models are called feedforward because the flow of information through them is always in the forward direction, and no feedback connections (i.e., connections in which the outputs of a model are fed back to it as input) are allowed. Generally, a sequential order must be held between the layers of the MLP network architecture. Like typical artificial neural networks, multilayer perceptron networks can also be used for such various purposes as prediction, classification, and clustering. In particular, when a large number of input variables are involved, or in cases where the nature of the input corresponds to an N-dimensional array, a deep multilayer network design needs to be employed.

Because the data sets we use in this study contain multiple multi-class categorical features (e.g., diagnosis code), each with many different potential values, employing multiple regression or other classic machine learning techniques requires a sizeable reduction in the dimensionality of those variables. This is true due to the *curse of dimensionality* as a result of numerous levels of these variables, which in

[3] Using a computer with a 2.6 GHz Core i7 CPU and 16 GB of RAM, preprocessing each data set takes approximately 50 to 60 minutes.

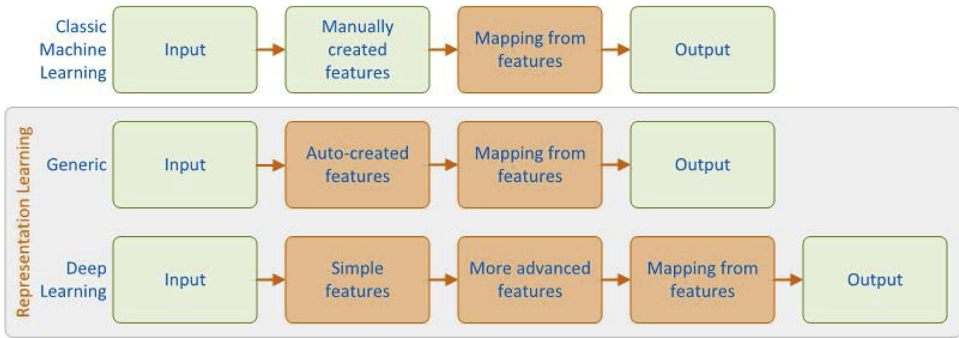


Fig. 4. Classic machine learning vs. deep learning methods.

turn, influence the statistical significance of the regression parameters and lead to inefficiency in other algorithms [53]. Various techniques are proposed in the machine learning literature to restrain the curse of dimensionality and avoid its consequences on data analysis using regression or classic machine learning methods [54–56]. With deep neural networks, however, prior research argues that the curse of dimensionality is not a serious problem [53,57,58].

In addition, reducing the dimensionality of multi-class categorical features (even if some of the more advanced machine learning methods are used to build predictive models) results in the loss of a significant amount of information and seriously affects the quality of the predictions. To address this issue, therefore, the current study employs an MLP deep learning network for predicting patients' length of hospital stay.

Essentially, the choice of an appropriate network architecture for deep learning is highly dictated by the type and nature of the data, as well as the way it is prepared for analysis. We chose MLP over two other popular deep learning architectures, namely, Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN). We did so because even though these architectures are very powerful in feature extraction and pattern recognition, each of them requires a different type of data to work efficiently. LSTM is basically an architecture designed for analyzing sequences where a meaningful order (either spatial or temporal) is associated with the data points. For instance Peng et al. [59] use an LSTM network with a time-series data set (i.e., data points have a temporal order) to predict electricity prices.

In contrast, the CNN architecture was originally devised for image processing; however, it is also applicable to data sets where each data point can be attributed to a matrix (2D) or a higher dimensional tensor, as opposed to a one-dimensional array. For example, Li et al. [60] use a CNN architecture to analyze the electrocardiogram (ECG) pictures of patients' heartbeats. Since the data sets that we use in this study are neither sequential nor high dimensional tensors, we choose to employ an MLP architecture to analyze our one-dimensional data points.

Because MLP networks work only with numerical inputs, we keep the numerical predictors intact, but apply one-hot encoding to all of the categorical variables included in each data set. As a result, each categorical predictor with m distinct levels is converted to $m-1$ binary variables, yielding a total of 59 numeric predictors in each data set. Moreover, to avoid any bias due to different measurement units, we perform a min-max normalization to represent the values of all variables in the [0, 1] range.

In order to obtain a nearly optimal network architecture in a systematic manner, we begin by experimenting with three different architectures including five, seven, and ten hidden layers (each with 128 neurons) and train the networks with the same parameter settings (i.e., the software defaults). Table 4 shows the accuracy measures of the three models on the test portion of the COPD data set. Given the better performance of the network with five hidden layers (probably due to the fast overfitting of the networks with more layers), we choose this architecture as the base and try to optimize its hyper parameters to improve its output.

In the next step, we perform a series of experiments to obtain a nearly optimal number for the neurons in each layer. To this end, we train four different network architectures, all including five hidden layers, but with 64, 128, 256, and 512 neurons in each layer. Fig. 5 displays the R^2 of each network on the test data set as we increase the width of the network (i.e., the number of neurons). It is clear that the 5-layer architecture with 256 neurons in each layer outperforms the other architectures. The figure suggests that, if the default learning parameters are used, increasing the width of the network up to 256 neurons in each layer improves the performance of the network. However, doing so beyond 256 neurons in each layer makes the network too dense, resulting in quick learning with respect to the patterns in the training data set (i.e., overfitting).

Now that we have obtained a rough idea of a decent architecture for our deep learning network, we gradually modify the network parameters based on the best practice guidelines to gain a nearly optimal training configuration. This involves a long process of slightly manipulating the learning rate, number of epochs, batch size, number of neurons (in single layers), and learning stoppage criteria. Ultimately, we construct a fully connected MLP network that includes a visible input layer with 59 neurons, five dense hidden layers (the first one with 128 and the others with 256 neurons each), and a single-neuron output layer. Fig. 6 displays the MLP network's architecture.

For each hidden layer, we use a Rectified Linear Units (ReLU) function as the layer's activation (transfer) function. According to the extant literature [61], ReLU is a very efficient activation function for deep learning applications because unlike other popular functions (such as the sigmoid or the hyperbolic tangent), it does not generally suffer from the *vanishing*

Table 4
Performance metrics of the initial architectures.

Model	R^2	MAE	MSE	RMSE
5 hidden layers	0.568	1.322	4.571	2.402
7 hidden layers	0.533	1.412	4.877	2.521
10 hidden layers	0.464	1.703	5.101	2.694

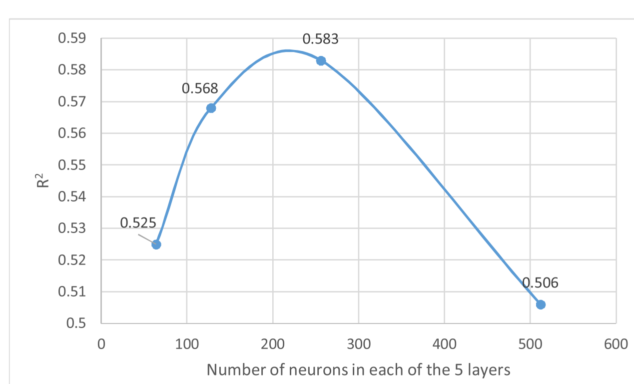


Fig. 5. Performance of a 5-layer architecture with a different number of neurons in each layer.

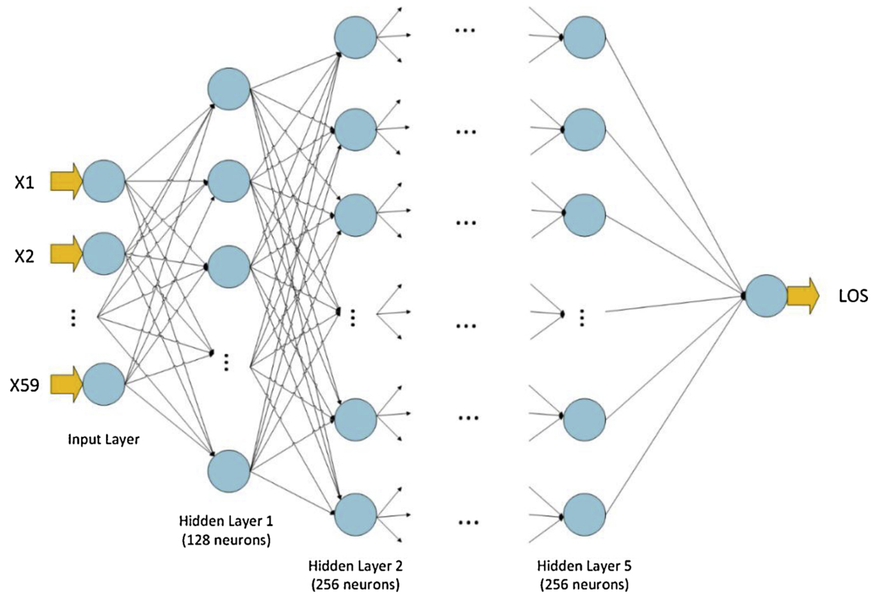


Fig. 6. The MLP network architecture.

gradient problem. This problem usually occurs when the derivative of the activation function for a given input is very small, leading to difficulties in updating the network weights through the gradient descent optimization process. Finally, since the network's output has to be a real number (i.e., LOS), we use a linear activation function for the output layer.

To implement the designed deep MLP network, we use Keras (https://keras.io), an open source Python library. Keras is an application programming interface with the capability of working on top of a number of popular deep learning frameworks, such as TensorFlow, which is the framework of choice in this effort. To be able to perform all of the model-building activities (i.e., data cleaning, variable transformation, model training, and evaluation) in the same environment, we use a Keras integration add-on within KNIME, a java-based, free, and open-source data analytics reporting and integration platform.

Having provided the details of our deep MLP network, we now turn our focus to the analytical evaluation of this model and to the contribution of our proposed data engineering methodology to the models' overall performance.

5. Analytical evaluation

To evaluate the contribution of our data engineering and model building processes to the prediction of patients' length of hospital stay, we use temporal validation along with the standard assessment metrics for predictive models [62]. We partition each data set into training and test subsets by using the earlier three quarters of the hospital encounters for training and the latter quarter for evaluation purposes. Summary statistics of the training and testing subsets for each of the two data sets are presented in Table 5.

We use the training partition of each data set to train the deep MLP network with the *mean absolute error (MAE)* as the loss function and the *ADAM* optimization algorithm [63] as the network optimizer. Within each training partition, we allocate 20 % of the encounters to a preliminary model validation with the aim of avoiding overfitting. The training procedure involves a maximum of 1000 epochs with a batch size of 200 admissions 273 steps per epoch for the COPD and 188 steps per epoch for the pneumonia data. In each case, to avoid overfitting, we stop the training process when no improvement is observed in the MAE of the validation subset in 20 consecutive epochs. The training procedure for each data set takes approximately 20 min on a powerful machine with four NVIDIA TITAN GPUs (used in parallel) and 64GB of memory.

Table 5
Demographics of the training and testing data.

Dataset	Variable	Statistic / levels	Training	Testing
COPD		Total admissions	68,216 (79 %)	18,122 (21 %)
	LOS	Mean (Std. Dev.)	4.68 (3.46)	4.47 (3.32)
	Age	Mean (std. Dev.)	56.71 (22.17)	55.75 (22.22)
	Gender	Male	40.3 %	39.8 %
		Female	59.7 %	60.2 %
	Race	White	73.3 %	72.5 %
		Black	20.0 %	20.3 %
		Other	6.7 %	7.2 %
Pneumonia		Total admissions	46,951 (74.3 %)	16,234 (25.7 %)
	LOS	Mean (Std. Dev.)	7.51 (6.57)	7.20 (6.25)
	Age	Mean (Std. Dev.)	61.22 (23.48)	61.72 (23.24)
	Gender	Male	50.1 %	49.7 %
		Female	49.9 %	50.3 %
	Race	White	73.5 %	73.8 %
		Black	20.4 %	19.1 %
		Other	6.1 %	7.1 %

Table 6 uses the results obtained from the testing partitions to summarize the performance of the best-trained MLP networks for each of the diseases. The table reports the overall R^2, MAE, mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) for the prediction of hospital LOS as a numerical variable. Furthermore, since prior studies (e.g., [3,12]) have

Table 6
Prediction results for LOS.

Dataset	Metric	Value
COPD	R^2	0.613
	MAE	1.239
	MSE	4.256
	RMSE	2.063
	MAPE	40.55 %
	± 2-day tolerance	86.05 %
	± 3-day tolerance	91.34 %
Pneumonia	R^2	0.655
	MAE	2.038
	MSE	13.501
	RMSE	3.675
	MAPE	43.81 %
	± 2 days tolerance	74.43 %
	± 3 days tolerance	84.58 %

Table 7
Comparison of the results with prior research.

Study	Condition	Number of Hospitals	Testing Method	Algorithm	Accuracy			
					R^2	MAE	± 2 days	± 3 days
Present study	COPD	182	Temporal	MLP	0.61	1.24	86 %	91 %
Present study	Pneumonia	202	Temporal	MLP	0.66	2.04	74 %	85 %
[38]	Heart failure	3[a]	Random	Regression Tree	0.79	1	NA	NA
[12]	Heart diseases	1	Random	ANN	NA	3.76	56 %	67 %
[39]	Knee Replacement	1	Random	Poisson Regression	NA	NA	56 %	77 %
[40]	Hip fracture	~40	Random	Random Forest	0.83	NA	NA	NA

[a] All three are Veteran Health Administration hospitals.

extended the definition of accuracy in the prediction of LOS to account for its high degree of variation, Table 6 also provides two other accuracy metrics. These metrics represent the percentage of hospital encounters for which LOS has been predicted within two or three days from its actual value. Therefore, an accuracy of ± 2 days, for instance, means that the absolute value of the difference between the predicted value of LOS and its observed value is *less than* two days[4] .

As Table 6 shows, the deep MLP networks built in this effort are able to explain more than 60 percent of the variation in the LOS of each data set. Numerical predictions of LOS are, on average, only 1.239 and 2.038 days away from their actual values in the COPD and pneumonia data sets, suggesting a remarkably high accuracy in predicting the exact length of hospital stay in days. With respect to the 2- and 3-day tolerance for LOS, our models obtain *very high* and *high* accuracy rates for the COPD and pneumonia data sets, respectively.

As pointed out earlier, a majority of the existing research is either descriptive or uses classification techniques to predict LOS as a binary or categorical variable (i.e., long vs. short stays with a pre-determined cutoff). Among the few others that employ regression-based or machine learning approaches to predict the actual value of LOS (see Table 7), the discrepancies between the index diseases hinder a direct comparison of their results. This study, however, addresses some of the shortcomings of prior efforts, and hence, provides a more robust approach for the prediction of the length of stay in hospitals. We classify these shortcomings into two areas: methodological and patient samples.

As can be seen in Table 7, all studies that predict LOS as a numerical variable use a random split to create the training and testing data partitions. This is potentially problematic because in practice, we should only use data from patients' past visits to predict an event of interest in the future. Put differently, a random split disregards the chronological order of events and uses information obtained from prospective encounters to predict the response variable in previous hospitalizations, leading to unrealistic and inflated prediction accuracies [64]. Using temporal evaluation is especially critical in the context of predicting the length of hospital stay at the time of admission, since we know nothing beyond the information collectable at the time of admission or extractable from patients' historical visits.

In addition to the aforementioned methodological issue, the studies enumerated in Table 7 are mostly based on a small number of hospitals. Except for the last study in the table[5], these studies are based on one or a few hospitals that serve certain demographic groups. Similarly, the number of encounters and unique patients in these studies are significantly smaller than those numbers in our data sets (more than 50,000 in each). As a final note regarding the samples used in the studies listed in Table 7, we can refer to the length of patients' stay in hospitals, which is usually fewer than ten days. Our data sets, however, as illustrated in Appendix B, include hospital encounters with much longer stays[6] . These facts suggest that the current effort carries out a more robust assessment of model performance and has greater generalizability compared to the extant research on the prediction of hospital LOS as a continuous variable.

In the next stage of our analytical evaluation, we focus on assessing the contribution of our data engineering methodology to the results obtained by the deep MLP network. We use variable importance as an objective measure to compare the predictive ability of input variables which, in turn, quantifies the extent to which the data engineering process contributes to the performance of the models. Unlike regression-based models, however, MLP networks do not provide an explicit metric to indicate the relative importance of predictors in predicting the outcome. Hence, we employ sensitivity analysis [43,64] to obtain the input variables' comparative importance. To this end, we drop the predictors one at a time, train the MLP network without these predictors, and note the change in the MAE of the predictions in the reduced model, as compared to the original network. Therefore, the variable without which the performance of the MLP network suffers the most (i.e., its MAE has the largest value) is identified as the most important variable. Next, we standardize the differences obtained across the whole set of predictors to create a "relative importance" index. Ultimately, we rank the predictors according to their index scores. Table 8 demonstrates the top ten predictors of LOS in each data set, along with their relative importance.

As shown, *pt_avg_los*, a variable calculated through our data engineering methodology, dwarfs the other variables in both data sets. In the COPD and pneumonia data, respectively, the MLP model's MAE increases by more than 1.0, and by almost 2.0 units, if we remove *pt_avg_los* from the input variables. The relative importance of this variable is more than three times than that of the second ranking variable (i.e., Age) in both data sets. Similarly, *pt_avg_readmission,* another derived variable representing the average time between consecutive readmissions of a patient across all of her past visits, ranks among the top four predictors of LOS in both data sets. In each of the data sets, a third variable created through the data engineering step is ranked in sixth place. These variables (i.e., *avg_readmission* and *avg_los*), which provide an overall comparison between the health status of a focal patient and that of the average patient in the population, have not been considered in earlier studies on the prediction of LOS in hospitals[7] . Hence, it is evident that the superior performance of the deep MLP model we developed and evaluated in this study is largely driven by our data engineering effort, which is based upon the value of augmenting electronic medical records with patients' historical information. In addition, the emergence of variables related to patients' medical histories among the most important variables advances the literature on the prediction of hospital

[4] The definition given by Tsai et al. [12] for the 2-day tolerance is that "the difference between the prediction of LOS and the actual LOS is less than 3 days."

[5] Elbattah and Molloy [40] did not report the number of different hospitals in the data they used. We surmise that the number is at most 40 by counting the number of hospitals subscribing to the Irish hip fracture database.

[6] This is true, even after removing the outliers from the data.

[7] In fact, only one study has considered medical history at the patient level in its predictive models. In that study, however, such information was readily available as independent variables in the data. Since most EMR data are stored at the transaction level, patients' medical histories can only be obtained by extensive data engineering. Thus, identifying the importance of medical history at the patient level is a theoretical contribution of this effort.

Table 8
Relative importance of the predictor variables.

COPD			Pneumonia		
Variable	MAE Change	Relative Importance	Variable	MAE Change	Relative Importance
pt_avg_los	1.032	1.000	pt_avg_los	1.948	1.000
Age	0.296	0.287	Age	0.598	0.307
Diagnosis	0.196	0.190	pt_avg_readmission	0.463	0.238
pt_avg_readmission	0.185	0.179	Bed_size	0.290	0.149
Has_pneumonia	0.159	0.154	Census_Division	0.268	0.138
avg_readmission	0.149	0.144	avg_los	0.218	0.112
Bed_Size	0.138	0.134	Diagnosis	0.191	0.098
Census_Division	0.132	0.127	Teaching_Facility	0.166	0.085
Has_heart_disease	0.119	0.115	Marital_Status	0.096	0.049
First_admission	0.071	0.069	Cath_Lab_Full_ind	0.083	0.043

LOS, especially in chronic diseases, by demonstrating the significant contribution of these variables to the overall accuracy of predictive models. This, in turn, encourages the collection of data concerning (or the aggregation of data from) patients' histories of health care use. Similarly, it underlines the importance of developing appropriate processes for sharing data between affiliated health care providers.

Another interesting finding from the list of important variables in Table 8 is the emergence of hospital size, measured by the number of beds, and the census division (i.e., where the hospital is located) among the topmost significant predictors. Several reasons may contribute to the high predictive power of these variables, such as differences in the quality of care provided by hospitals, lifestyle of the patients, or their socioeconomic status. Besides these factors, patients' exact diagnosis (i.e., the type of pneumonia or any of the specific diseases collectively referred to with the umbrella term COPD) is also important in determining their length of stay. Comorbid conditions (e.g., whether a patient is also diagnosed with chronic heart disease or diabetes) and whether the hospital is affiliated with a university are among the other important predictors for patients' LOS in the COPD and pneumonia data sets.

6. Discussion and conclusion

A diverse set of factors, related to hospitals' finances, customers, employees, and communities, are the forces behind hospitals' drive toward sustainable operations. It is estimated that the U.S. health care industry could save as much as $15 billion over the next decade simply by implementing more sustainable practices. Individual hospitals can save millions of dollars through waste reduction efforts, energy efficiency initiatives, and environmentally responsible purchasing [65]. These gains, in turn, would have a positive and measurable effect on the health of local communities, employees' quality of working life, and patients' satisfaction. One way hospitals could realize these benefits, make better use of their resources, and provide an optimal course of treatment to their patients is via predicting patients' length of hospital stay at the time of admission.

In this paper, we employed a data analytic approach to develop and test a deep neural network to predict patients' length of hospital stay. Our effort not only addressed the shortcomings of previous studies (i.e., ignoring patients' previous admissions, lack of temporal evaluation, and failing to predict LOS as a continuous variable), but also resulted in models with remarkably high accuracies. Furthermore, we showed how our data engineering method, which extracts patients' histories and appends them as new variables to electronic medical records, contributed the most to the models' accurate predictions. While it is widely known that a majority of the time involved in developing data analytic models is allocated to data preparation rather than to building the models, how this task is performed bears a significant impact on the quality of the outcomes. Therefore, besides the accurate prediction of patients' LOS, this paper contributes to the literature on sustainable health care and data analytic clinical decision support systems (CDSS) in two dimensions: theory and methodology. In the theoretical dimension, we introduced a new paradigm for the prediction of LOS in hospitals by verifying the importance of patients' previous admissions. In the methodological dimension, we demonstrated the importance of temporal evaluation and data engineering.

Temporal evaluation, in which a portion of earlier encounters is used for model training and the remaining latter portion is used for validation, is a crucial factor in the prediction of events with a chronological order. As a result, using a random split that disregards the order of events may lead to unrealistic and inflated predictions. Our use of temporal evaluation, together with the exclusion of all variables whose values are unknown at the time of admission, illustrates a more realistic and practical approach to building CDSS. Moreover, our data engineering methodology highlighted the importance of considering medical data as a collection of interrelated (rather than transactional) records. The fact that the variables created by extracting information from patients' prior visits turned out to be the most important predictors of LOS suggests that the current paradigm, in which only variables related to a patient's current visit are included in the analyses, needs to be revised. In addition, the emergence of variables related to patients' previous admissions among the most important predictors of LOS suggests that information sharing among affiliated hospitals should be encouraged. In doing so, more historical medical records (especially from recent years) can be easily traced and augmented for better prediction results. Regarding the size of electronic medical records and the sophistication of aggregating and processing such data, another implication of our study is that it underlines the urgency for hospital systems to invest in their analytical capabilities. By doing so, they can make more data-driven decisions in regard to various aspects of providing health care services to their patients. Based on our discussions, such endeavors will not only furnish hospitals with financial benefits, but will also result in improved health outcomes, enhanced services to patients (e.g., better treatment planning and reduced bills), and ultimately, more sustainable operations.

As mentioned earlier, the resource-intensive computations required to aggregate data from various sources, such as the extraction and addition of patients' historical information and the time-consuming adjustments required to assure the veracity of the health data, are the main big data challenges in developing CDSS. Regarding all of these intricacies, deep learning, with its automatic feature extraction, is building its way into the world of developing clinical decision aid tools. For this reason, we believe that our development and deployment of a CDSS with deep learning not only adds to the scholarly literature, but can also spur interest among practitioners in a number of ways. First, our use of patients' historical data to improve the prediction of hospital LOS among chronic patients accentuates the need to invest in appropriate infrastructures that not only control the quality and veracity of medical data, but also enable easier sharing of such data within a health care system or between partnering organizations.

Second, while demonstrating the utility of data analytic techniques in improving health outcomes, this study provides an illustrative example of how data engineering and analytics can be employed to realize such benefits. Third, this study lists the most important predictors of LOS in two common chronic diseases, thereby advising health care managers and practitioners of the factors that should be given more weight when a patient's LOS is predicted at the time of admission. Finally, for managers and practitioners who are interested in applications of data analytic techniques to health care, it introduces the concept of the temporal evaluation of predictive models, and suggests when a random partitioning of training and test data should be avoided.

We caution against the assumption that our approach would result in similarly accurate predictions for other diseases. The complex nature of different diseases, how they interact with comorbid conditions, treatments, or drugs, or how they are affected by such external factors as patients' lifestyle or socioeconomic status, may hamper such generalizations (unless a large number of studies find similar results for other chronic diseases). Thus, investigating whether the information obtained from patients' prior hospitalizations can be used to improve the prediction of LOS in other diseases remains an open avenue for future research. Another possibility for future research is investigating the optimal timespan for extracting patients' historical information to obtain the best results in terms of both the predictions and time complexity of the computations.

A limitation of our research is that it used data sets that had fewer variables compared to certain prior studies. While obtaining high accuracies with fewer variables may be another indication of the paper's contributions, it may have been possible to further improve our decision support tool by adding more variables. Finally, the paucity of prior research on the prediction of LOS as a numerical variable limited our ability to compare the performance of our models with previous studies. However, as we discussed in the paper, since we employed a more robust approach to the development and evaluation of the predictive models, our findings should still be informative to researchers and practitioners.

CRediT authorship contribution statement

Hamed M. Zolbanin: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Project administration. **Behrooz Davazdahemami:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing. **Dursun Delen:** Resources, Visualization, Writing - review & editing, Supervision. **Amir Hassan Zadeh:** Resources, Writing - review & editing.

Acknowledgement

This study was conducted with the data provided by, and the support from, the Center for Health Systems Innovation (CHSI) at Oklahoma State University (OSU) and the Cerner Corporation. The contents of this work are solely the responsibility of the authors and do not necessarily represent the official views of CHSI, OSU or the Cerner Corporation.

Appendix A. Summary of Prior Studies on the Prediction of LOS in Hospitals

Study	Condition	Setup	Best Performing Model	Operationalization of LOS	Important Factors	Gap
[9]	Stroke	Predictive	Bayesian network	Nominal	NIH Stroke Score Mode of arrival Pneumonia Sensory changes Age group	LOS operationalization Random split Medical history
[6]	Multiple	Predictive	Regression Random Forest	Binary	Age / Insurance / Status Reason for the visit Day of admission Discharge location	LOS operationalization Random split Medical history
[66]	Psychiatry	Descriptive	ANOVA and Linear Regression	Numerical	Active duty military status Race /Paranoia Violence / Suicidality Physical restraint Personality disorder (at discharge)	Descriptive Medical history Use of data obtained during the hospitalization
[39]	Knee Replacement	Predictive	Poisson Regression	Numerical	Age /Gender Day of admission Discharge location	Random split Medical history
[31]	Open AAA surgery	Descriptive	Logistic Regression	Binary	Disease factors Race / COPD / Age Admission source	Descriptive Medical history
[32]	Multiple (Emergency Department)	Descriptive	Survival Analysis	Numerical	Age / Gender Shift of admission Admitting doctor Day of admission Lab procedures	Descriptive Medical history
[67]	Type 2 Diabetes	Descriptive	Linear Regression	Numerical	Gender / Age group Insurance / Occupation Type of admission Comorbidities	Descriptive Medical history
[15]	Trauma	Descriptive	Binomial Regression	Numerical	Insurance /Age / Prior visits Admission month Employee Collaboration Procedures	Descriptive Medical history
[68]	Psychiatry	Descriptive	ANOVA & Linear Regression	Numerical	Age / Living alone Psychiatric symptoms Social behavior score	Descriptive Medical history
[27]	Lung Resection	Predictive	Logistic Regression	Binary	Age / COPD / Transfusion Sodium levels Open thoracotomy Comorbidities Return to the operation room	LOS operationalization Random split Medical history

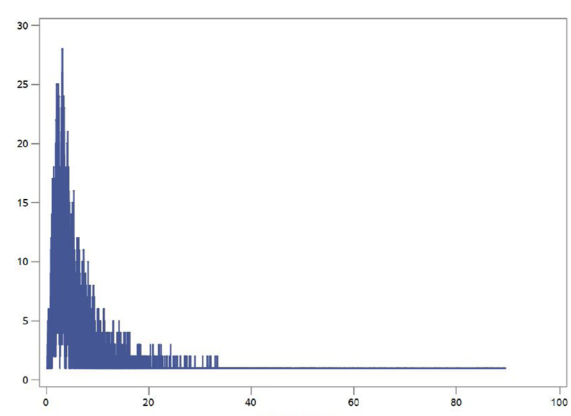
[40]	Hip Fracture	Predictive	Random Forest	Numerical	Admission type & source Age / Gender Discharge status Fracture type Fragility history Residence area Pre-fracture mobility	Random split Medical history
[33]	Hip Fracture	Descriptive	Linear Regression	Numerical	Fitness of patients for surgery Gender / Emergency Smoking Status	Descriptive Medical history
[69]	Hip Replacement	Descriptive	Negative Binomial Regression	Numerical	Age group / Gender Comorbidity Index Diagnosis related variables	Descriptive Medical history
[70]	Coronary Artery Bypass Surgery	Descriptive	Linear Regression	Numerical	Age / Race / Gender Socioeconomic status Comorbidity Disease Severity	Descriptive Medical history Use of data obtained during the hospitalization
[28]	Trauma	Predictive	Artificial Neural Networks	Binary	Trauma and Injury Severity Score	LOS operationalization Random split Medical history Large gap between specificity and sensitivity
[18]	Coronary Artery	Predictive	Support Vector Machines	Nominal	Blood pressure Age / Smoking status Anticoagulant drugs Ejection fraction Nitrate drugs	LOS operationalization Random split Medical history Reports performance on the training data
[29]	Multiple (Emergency Department)	Predictive	Artificial Neural Networks	Binary	Age / Gender History of falls Acute organ failure Reason for visit Lacking home help Living at home	LOS operationalization Random split Medical history Large gap between specificity and sensitivity
[71]	Pneumonia	Descriptive	Linear Regression	Numerical	Age / Gender Diagnosis Related Group Severity	Descriptive Medical History
[34]	Multiple	Descriptive	Linear & Logistic Regression	Numerical	Primary condition Comorbidity score Physiology score Age / Admission shift Day of admission	Descriptive Medical history
[72]	Multiple	Descriptive	Linear Regression	Numerical	Sociodemographic Admission source Insurance Diagnosis related group Specialty of the doctor at time of discharge	Descriptive Medical History
[73]	Acute Chest Pain	Descriptive	Linear Regression	Numerical	Comorbidity Index	Descriptive Medical history
[74]	Psychiatry	Descriptive	Linear Regression	Numerical	Age/ Gender Diagnostic categories	Descriptive Medical history
[75]	Multiple	Predictive	Decision Tree	Numerical	Age / Gender Lab data Major Diagnostic Categories	Random split Medical history Use of data obtained during the hospitalization
[76]	Inguinal Hernia	Descriptive	Poisson Regression	Numerical	Age / Gender Number of diagnoses Number of procedure	Descriptive Medical history
[77]	Multiple	Descriptive	Linear Regression	Numerical	Age / Gender / Race Chronic Diseases Score Comorbidity Index	Descriptive Medical history
[78]	Stroke	Descriptive	Negative Binomial Regression	Numerical	Age / Gender Diagnosis related groups Comorbidity index	Descriptive Medical history
[79]	Psychiatry	Descriptive	Hierarchical Regression	Numerical	Age Diagnosis related groups Patient, service, and area related variables	Descriptive Medical history
[30]	COPD	Predictive	Linear & Logistic Regression	Binary	Number of last year admissions (≥ 3 or < 3) Admitted in last month? Heart disease Source of admission	LOS operationalization Random split Medical history
[80]	Chronic Disability	Descriptive	Linear Regression	Numerical	Age / Comorbidity index	Descriptive Medical history
[4]	Cardiac Surgery	Predictive	Artificial Neural Networks	Binary	Age / Gender Heart rate Comorbidity Operative priority Prior myocardial infarction Respiratory rate	LOS operationalization Random split Medical history

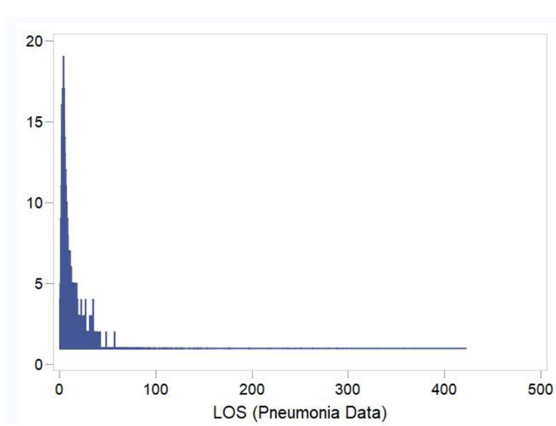
[35]	COPD	Descriptive	Logistic Regression	Binary	Smoking status Comorbidity score Body mass index Previous admissions Treatment factors	LOS operationalization Random split Medical history
[81]	Multiple	Descriptive	Linear Regression	Numerical	Age Functional status Number of therapy referrals	Descriptive Medical history Use of data obtained during the hospitalization
[36]	Stroke	Descriptive	Logistic Regression	Binary	Age / Previous stroke Premorbid disability Comorbid heart diseases Fever in first 5 days Stroke in progression	LOS operationalization Random split Medical history
[12]	Multiple (Cardiology)	Predictive	Artificial Neural Networks	Numerical	Gender / Location Main diagnosis Comorbidity Interventions	Random split Medical history
[38]	Heart Failure	Predictive	Regression Tree	Numerical	Number of admissions in the past quarters Number of all hospital stays (last month & year) Total LOS Number of admissions	Random Split Medical history is only at the patient level
[82]	Psychiatry	Descriptive	ANCOVA	Numerical	Severity of disorder Diagnosis related group	Descriptive Medical history
[37]	Heart Failure	Descriptive	Linear Regression	Nominal	Comorbid conditions Heart rate / Insurance Day of admission / Smoking status Age / Gender / Race	Descriptive Medical history
[19]	COPD	Predictive	Logistic Regression	Binary	Day of admission Comorbid conditions Serum albumin level	Random split Medical history

[a]This column only lists some of the important factors in each study. Except for one, however, none of the studies included patients' medical histories in their models. See the entry for Turgeman et al. [38] for more details.

[b]Abdominal Aortic Aneurysm.

Appendix B. Distribution of LOS in the Two Data Sets





Appendix C. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.im.2020.103282.

References

[1] Becker's Healthcare, How Predicting Patient Length of Stay Enables Hospitals to Save Millions, Retrieved January 7, 2019, from (2018) https://go.beckershospitalreview.com/how-predicting-patient-length-of-stay-enables-hospitals-to-save-millions.

[2] C.M. Kozma, M. Dickson, M.K. Raut, S. Mody, A.C. Fisher, J.R. Schein, J.I. Mackowiak, Economic benefit of a 1-day reduction in hospital stay for community-acquired pneumonia (CAP), J. Med. Econ. 13 (4) (2010) 719–727, https://doi.org/10.3111/13696998.2010.536350.

[3] Intel, Predictive Analytics Help Hospital Predict Patient Length-of-Stay, Retrieved January 7, 2019, from (2015) https://www.intel.com/content/www/us/en/healthcare-it/solutions/documents/large-hospital-group-allocate-resources-by-predicting-length-of-stay-study.html.

[4] M. Rowan, T. Ryan, F. Hegarty, N. O'Hare, The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors, Artif. Intell. Med. 40 (3) (2007) 211–221, https://doi.org/10.1016/j.artmed.2007.04.005.

[5] L. Turgeman, J.H. May, A mixed-ensemble model for hospital readmission, Artif. Intell. Med. 72 (2016) 72–82, https://doi.org/10.1016/j.artmed.2016.08.005.

[6] S. Barnes, E. Hamrock, M. Toerper, S. Siddiqui, S. Levin, Real-time prediction of inpatient length of stay for discharge prioritization, J. Am. Med. Inform. Assoc. 23 (e1) (2016) e2–e10, https://doi.org/10.1093/jamia/ocv106.

[7] C.-L. Lin, P.-H. Lin, L.-W. Chou, S.-J. Lan, N.-H. Meng, S.-F. Lo, H.-D.I. Wu, Model-based prediction of length of stay for rehabilitating stroke patients, J. Formos. Med. Assoc. 108 (8) (2009) 653–662 Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19666353.

[8] V. Nerminathan, W. Adlan, A. Nerminathan, Proceedings book of ICEFMO, 2013, Malaysia handbook on the economic, finance and management outlooks measuring training effectiveness: evidence from Malaysia, Int. J. Manag. Sustain. 3 (2) (2014) 51–61 Retrieved from http://www.conscientiabeam.com/ebooks/ICEFMO-131-455-463.pdf.

[9] A.R. Al Taleb, M. Hoque, A. Hasanat, M.B. Khan, Application of data mining techniques to predict length of stay of stroke patients, 2017 International Conference on Informatics, Health & Technology (ICIHT), 2017, pp. 1–5, , https://doi.org/10.1109/ICIHT.2017.7899004.

[10] L. Lella, I. Licata, Prediction of length of Hospital stay using a growing neural gas model, Proceedings of the 8th International Multi-Conferences on Complexity, Informatics and Cybernetics, 2017 Retrieved from http://www.iiis.org/CDs2017/CD2017Spring/papers/ZA254QL.pdf.

[11] M. Marimuthu, H. Paulose, Emergence of sustainability based approaches in healthcare: expanding research and practice, Procedia - Soc. Behav. Sci. 224 (2016) 554–561, https://doi.org/10.1016/J.SBSPRO.2016.05.437.

[12] P.-F.(Jennifer) Tsai, P.-C. Chen, Y.-Y. Chen, H.-Y. Song, H.-M. Lin, F.-M. Lin, Q.-P. Huang, Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network, J. Healthc. Eng. 2016 (2016) 1–11, https://doi.org/10.1155/2016/7035463.

[13] S. Walczak, W.E. Pofahl, R.J. Scorpio, A decision support tool for allocating hospital bed resources and determining required acuity of care, Decis. Support Syst. 34 (4) (2003) 445–456, https://doi.org/10.1016/S0167-9236(02)00071-4.

[14] H. Alosh, D. Li, L.H. Riley, R.L. Skolasky, Health care burden of anterior cervical spine surgery, J. Spinal Disord. Tech. 28 (1) (2015) 5–11, https://doi.org/10.1097/BSD.0000000000000001.

[15] Y. Chen, M.B. Patel, C.D. McNaughton, B.A. Malin, Interaction patterns of trauma providers are associated with length of stay, J. Am. Med. Inform. Assoc. 25 (7) (2018) 790–799, https://doi.org/10.1093/jamia/ocy009.

[16] C. DeRienzo, J.A. Kohler, E. Lada, P. Meanor, D. Tanaka, Demonstrating the relationships of length of stay, cost and clinical outcomes in a simulated NICU, J. Perinatol. 36 (12) (2016) 1128–1131, https://doi.org/10.1038/jp.2016.128.

[17] A. Riascos, N. Serna, Predicting Annual Length-Of-Stay and Its Impact on Health Vol. 69 (2017), pp. 27–34 Retrieved from http://proceedings.mlr.press/v69/riascos17a.html.

[18] P.R. Hachesu, M. Ahmadi, S. Alizadeh, F. Sadoughi, Use of data mining techniques to determine and predict length of stay of cardiac patients, Healthc. Inform. Res. 19 (2) (2013) 121–129, https://doi.org/10.4258/hir.2013.19.2.121.

[19] Ying Wang, K. Stavem, F. Dahl, S. Humerfelt, T. Haugen, Factors associated with a prolonged length of stay after acute exacerbation of chronic obstructive pulmonary disease (AECOPD), Int. J. Chron. Obstruct. Pulmon. Dis. 9 (2014) 99, https://doi.org/10.2147/COPD.S51467.

[20] R. Nambiar, R. Bhardwaj, A. Sethi, R. Vargheese, A look at Challenges and Opportunities of Big Data Analytics in Healthcare, 2013 IEEE International Conference on Big Data, 2013, pp. 17–22, , https://doi.org/10.1109/BigData.2013.6691753.

[21] M.J. Ward, K.A. Marsolo, C.M. Froehle, Applications of business analytics in healthcare, Bus. Horiz. 57 (5) (2014) 571–582, https://doi.org/10.1016/j.bushor.2014.06.003.

[22] Yichuan Wang, L. Kung, T.A. Byrd, Big data analytics: understanding its capabilities and potential benefits for healthcare organizations, Technol. Forecast. Soc. Change 126 (2018) 3–13, https://doi.org/10.1016/j.techfore.2015.12.019.

[23] Yichuan Wang, L. Kung, W.Y.C. Wang, C.G. Cegielski, An integrated big data analytics-enabled transformation model: Application to health care, Inf. Manag. 55 (1) (2018) 64–79, https://doi.org/10.1016/j.im.2017.04.001.

[24] Knowledge@Wharton, The Promise of Big Data: Revolutionizing Health Care Sustainability - Knowledge@Wharton, Retrieved January 9, 2019, from (2016) http://knowledge.wharton.upenn.edu/article/promise-big-data-information-can-revolutionize-health-care-sustainability/.

[25] HPOE, Environmental sustainability in hospitals: the value of efficiency, Hospitals in Pursuit of Excellence (May) (2014) 1–34.

[26] E. Pantzartzis, F.T. Edum-Fotwe, A.D.F. Price, Sustainable healthcare facilities: reconciling bed capacity and local needs, Int. J. Sustain. Built Environ. 6 (1) (2017) 54–68, https://doi.org/10.1016/J.IJSBE.2017.01.003.

[27] M.R. DeLuzio, H.B. Keshava, Z. Wang, D.J. Boffa, F.C. Detterbeck, A.W. Kim, A model for predicting prolonged length of stay in patients undergoing anatomical lung resection: a National Surgical Quality Improvement Program (NSQIP) database study, Interact. Cardiovasc. Thorac. Surg. 23 (2) (2016) 208–215, https://doi.org/10.1093/icvts/ivw090.

[28] C. Gholipour, F. Rahim, A. Fakhree, B. Ziapour, Using an artificial neural networks (ANNs) model for prediction of intensive care unit (ICU) outcome and length of stay at hospital in traumatic patients, J. Clin. Diagn. Res. 9 (4) (2015) OC19–OC23, https://doi.org/10.7860/JCDR/2015/9467.5828.

[29] C.P. Launay, H. Rivière, A. Kabeshova, O. Beauchet, Predicting prolonged length of hospital stay in older emergency department users: use of a novel analysis method, the Artificial Neural Network, Eur. J. Intern. Med. 26 (7) (2015) 478–482, https://doi.org/10.1016/j.ejim.2015.06.002.

[30] J.M. Quintana, A. Unzurrunzaga, S. Garcia-Gutierrez, N. Gonzalez, I. Lafuente, M. Bare, et al., Predictors of hospital length of stay in patients with exacerbations of COPD: a cohort study, J. Gen. Intern. Med. 30 (6) (2015) 824–831, https://doi.org/10.1007/s11606-014-3129-x.

[31] S. Casillas-Berumen, F.A. Rojas-Miguez, A. Farber, S. Komshian, J.A. Kalish, D. Rybin, et al., Patient and aneurysm characteristics predicting prolonged length of stay after elective open AAA repair in the endovascular era, Vasc. Endovascular Surg. 52 (1) (2017) 5–10, https://doi.org/10.1177/1538574417739747.

[32] C.-H. Chaou, H.-H. Chen, S.-H. Chang, P. Tang, S.-L. Pan, A.M.-F. Yen, T.-F. Chiu, Predicting length of stay among patients discharged from the emergency department—using an accelerated failure time model, PLoS One 12 (1) (2017) e0165756, https://doi.org/10.1371/journal.pone.0165756.

[33] A.E. Garcia, J.V. Bonnaig, Z.T. Yoneda, J.E. Richards, J.M. Ehrenfeld, W.T. Obremskey, et al., Patient variables which may predict length of stay and hospital costs in elderly patients with hip fracture, J. Orthop. Trauma 26 (11) (2012) 620–623, https://doi.org/10.1097/BOT.0b013e3182695416.

[34] V. Liu, P. Kipnis, M.K. Gould, G.J. Escobar, Length of stay predictions, Med. Care 48 (8) (2010) 739–744, https://doi.org/10.1097/MLR.0b013e3181e359f3.

[35] M. Ruparel, J.L. López-Campos, A. Castro-Acosta, S. Hartl, F. Pozo-Rodriguez, C.M. Roberts, Understanding variation in length of hospital stay for COPD exacerbation: european COPD audit, ERJ Open Res. 2 (1) (2016), https://doi.org/10.1183/23120541.00034-2015.

[36] N. Spratt, Y. Wang, C. Levi, K. Ng, M. Evans, J. Fisher, A prospective study of predictors of prolonged hospital stay and disability after stroke, J. Clin. Neurosci. 10 (6) (2003) 665–669 Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14592613.

[37] D.J. Whellan, X. Zhao, A.F. Hernandez, L. Liang, E.D. Peterson, D.L. Bhatt, et al., Predictors of hospital length of stay in heart failure: findings from get with the guidelines, J. Card. Fail. 17 (8) (2011) 649–656, https://doi.org/10.1016/j.cardfail.2011.04.005.

[38] L. Turgeman, J.H. May, R. Sciulli, Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission, Expert Syst. Appl. 78 (2017) 376–385, https://doi.org/10.1016/J.ESWA.2017.02.023.

[39] E.M. Carter, H.W. Potts, Predicting length of stay from an electronic patient record system: a primary total knee replacement example, BMC Med. Inform. Decis. Mak. 14 (1) (2014) 26, https://doi.org/10.1186/1472-6947-14-26.

[40] M. Elbattah, O. Molloy, Using Machine Learning to Predict Length of Stay and Discharge Destination for Hip-Fracture Patients, (2018), https://doi.org/10.1007/978-3-319-56994-9_15.

[41] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, https://doi.org/10.1038/nature14539.

[42] M.R. Cowie, J.I. Blomster, L.H. Curtis, S. Duclaux, I. Ford, F. Fritz, et al., Electronic health records to facilitate clinical research, Clin. Res. Cardiol. 106 (1) (2017) 1–9.

[43] H.M. Zolbanin, D. Delen, Processing electronic medical records to improve predictive analytics outcomes for hospital readmissions, Decis. Support Syst. 112 (2018) 98–110, https://doi.org/10.1016/J.DSS.2018.06.010.

[44] M. Herland, T.M. Khoshgoftaar, R. Wald, A review of data mining using big data in health informatics, J. Big Data 1 (1) (2014) 2, https://doi.org/10.1186/2196-1115-1-2.

[45] A. Nowiński, D. Kamiński, D. Korzybski, A. Stokłosa, D. Górecka, [The impact of comorbidities on the length of hospital treatment in patients with chronic obstructive pulmonary disease], Pneumonol. Alergol. Pol. 79 (6) (2011) 388–396 Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22028117.

[46] C. van Walraven, J. Wong, A.J. Forster, S. Hawken, Predicting post-discharge death or readmission: deterioration of model performance in population having multiple admissions per patient, J. Eval. Clin. Pract. 19 (6) (2013) 1012–1018, https://doi.org/10.1111/jep.12012.
[47] H. Wang, C. Johnson, R.D. Robinson, V.A. Nejtek, C.D. Schrader, J. Leuck, et al., Roles of disease severity and post-discharge outpatient visits as predictors of hospital readmissions, BMC Health Serv. Res. 16 (1) (2016) 564, https://doi.org/10.1186/s12913-016-1814-7.
[48] UNC Health Care Clinical Documentation Handbook, UNC Health Care Clinical Documentation Handbook, Retrieved from (2012) https://acdis.org/system/files/resources/287634.pdf.
[49] G. Rodriguez, Lecture Notes on Generalized Linear Models, Retrieved from (2007) https://data.princeton.edu/wws509/notes.
[50] D. Bhalla, Impute Missing Values With Decision Tree, Retrieved June 16, 2019, from (2015) https://www.listendata.com/2015/08/impute-missing-values-with-decision-tree.html.
[51] SAS Institute Inc, Data Mining Using SAS® Enterprise MinerTM: A Case Study Approach, p. 39). p. 39. Retrieved from, third edition, SAS Institute Inc., Cary, NC, 2013https://support.sas.com/documentation/cdl/en/emcs/66392/PDF/default/emcs.pdf.
[52] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.
[53] E. Keogh, A. Mueen, Curse of dimensionality, Encyclopedia of Machine Learning and Data Mining, Springer, 2017, pp. 314–315.
[54] F. Bach, Breaking the curse of dimensionality with convex neural networks, J. Mach. Learn. Res. 18 (19) (2017) 1–53.
[55] M.A. Bessa, R. Bostanabad, Z. Liu, A. Hu, D.W. Apley, C. Brinson, et al., A framework for data-driven analysis of materials under uncertainty: Countering the curse of dimensionality, Comput. Methods Appl. Mech. Eng. 320 (2017) 633–667.
[56] P. Rebeschini, R. Van Handel, Can local particle filters beat the curse of dimensionality? Ann. Appl. Probab. 25 (5) (2015) 2809–2866.
[57] J. Han, A. Jentzen, Overcoming the curse of dimensionality: solving high-dimensional partial differential equations using deep learning, ArXiv Preprint (2017) ArXiv:1707.02568.
[58] T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, Q. Liao, Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review, Int. J. Autom. Comput. 14 (5) (2017) 503–519.
[59] L. Peng, S. Liu, R. Liu, L. Wang, Effective long short-term memory with differential evolution algorithm for electricity price prediction, Energy 162 (2018) 1301–1314.
[60] Y. Li, Y. Pang, J. Wang, X. Li, Patient-specific ECG classification by deeper CNN from generic to dedicated, Neurocomputing 314 (2018) 336–346.
[61] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier Nonlinearities Improve Neural Network Acoustic Models, Retrieved from (2013) https://pdfs.semanticscholar.org/367f/2c63a6f6a10b3b64b8729d601e69337ee3cc.pdf.
[62] G. Shmueli, O.R. Koppius, Predictive analytics in information systems research, Mis Q. 35 (2011) 553–572, https://doi.org/10.2307/23042796.
[63] D.P. Kingma, J. Lei Ba, Adam: A Method for stochastic optimization, Retrieved from (2014) https://arxiv.org/pdf/1412.6980.pdf.
[64] B. Davazdahemami, D. Delen, A chronological pharmacovigilance network analytics approach for predicting adverse drug events, J. Am. Med. Inform. Assoc. (2018), https://doi.org/10.1093/jamia/ocy097.
[65] B. Wagner, Benefits of Sustainability in Healthcare Facilities - Facilities Management Insights, Retrieved from (2017) https://www.facilitiesnet.com/healthcarefacilities/article/Benefits-of-Sustainability-in-Healthcare-Facilities—17241.
[66] I.P. Brock III, G.R. Brown, Psychiatric length of stay determinants in a military medical center, Gen. Hosp. Psychiatry 15 (6) (1993) 392–398.
[67] D. Chen, S. Liu, X. Tan, Q. Zhao, Assessment of hospital length of stay and direct costs of type 2 diabetes in Hubei Province, China, BMC Health Serv. Res. 17 (1) (2017) 199, https://doi.org/10.1186/S12913-017-2140-4.
[68] F. Creed, B. Tomenson, P. Anthony, M. Tramner, Predicting length of stay in psychiatry, Psychol. Med. 27 (4) (1997) 961–966.
[69] A. Geissler, D. Scheller-Kreinsen, W. Quentin, Do diagnosis-related groups appropriately explain variations in costs and length of stay of hip replacement? A comparative assessment of DRG systems across 10 European countries, Health Econ. 21 (2012) 103–115.
[70] W.A. Ghali, R.E. Hall, A.S. Ash, M.A. Moskowitz, Identifying pre-and postoperative predictors of cost and length of stay for coronary artery bypass surgery, Am. J. Med. Qual. 14 (6) (1999) 248–254.
[71] L.I. Iezzoni, M. Shwartz, A.S. Ash, Y.D. Mackiernan, Does severity explain differences in hospital length of stay for pneumonia patients? J. Health Serv. Res. Policy 1 (2) (1996) 65–76.
[72] Y. Liu, M. Phillips, J. Codde, Factors influencing patients' length of stay, Aust. Health Rev. 24 (2) (2001) 63–70.
[73] K. Matsui, L. Goldman, P.A. Johnson, K.M. Kuntz, E.F. Cook, T.H. Lee, Comorbidity as a correlate of length of stay for hospitalized patients with acute chest pain, J. Gen. Intern. Med. 11 (5) (1996) 262–268.
[74] P. McCrone, M. Phelan, Diagnosis and length of psychiatric in-patient stay, Psychol. Med. 24 (4) (1994) 1025–1030.
[75] B. Mozes, M.J. Easterling, L.B. Sheiner, K.L. Melmon, R. Kline, E.S. Goldman, A.N. Brown, Case-mix adjustment using objective measures of severity: the case for laboratory data, Health Serv. Res. 28 (6) (1994) 689.
[76] J. O'Reilly, L. Serdén, M. Talbäck, B. McCarthy, Performance of 10 European DRG systems in explaining variation in resource utilisation in inguinal hernia repair, Health Econ. 21 (2012) 89–101.
[77] J.P. Parker, J.S. McCombs, E.A. Graddy, Can pharmacy data improve prediction of hospital outcomes? Comparisons with a diagnosis-based comorbidity measure, Med. Care (2003) 407–419.
[78] M. Peltola, E. Group, Patient classification and hospital costs of care for stroke in 10 European countries, Health Econ. 21 (2012) 129–140.
[79] R. Pertile, V. Donisi, L. Grigoletti, A. Angelozzi, G. Zamengo, G. Zulian, F. Amaddeo, DRGs and other patient-, service-and area-level factors influencing length of stay in acute psychiatric wards: the Veneto Region experience, Soc. Psychiatry Psychiatr. Epidemiol. 46 (7) (2011) 651–660.
[80] P.A. Rochon, J.N. Katz, L.A. Morrow, R. McGlinchey-Berroth, M.M. Ahlquist, M. Sarkarati, K.L. Minaker, Comorbid illness is associated with survival and length of hospital stay in patients with chronic disability: a prospective comparison of three comorbidity indices, Med. Care (1996) 1093–1101.
[81] S. Sahadevan, A. Earnest, Y.L. Koh, K.M. Lee, C.H. Soh, Y.Y. Ding, Improving the diagnosis related grouping model's ability to explain length of stay of elderly medical inpatients by incorporating function-linked variables, Ann. Acad. Med. Singap 33 (5) (2004) 614–622.
[82] I. Warnke, W. Rossler, Length of stay by ICD-based diagnostic groups as basis for the remuneration of psychiatric inpatient care in Switzerland? Swiss Med. 138 (35) (2008) 520.

Hamed M. Zolbanin is an assistant professor of information systems at the University of Dayton. Prior to this position, he served as the director of the business analytics program at Ball State University. He had several years of professional experience as an IT engineer prior to receiving his Ph.D. in Management Science and Information Systems from Oklahoma State University. His research has been published in such journal as *Decision Support Systems*, *Information Systems Frontiers*, and the *Journal of Business Research*. His main research interests are health care analytics, online reviews, sharing economy, and digital entrepreneurship.

Behrooz Davazdahemami is an Assistant Professor of Information Technology and Supply Chain Management (IT&SCM) at University of Wisconsin-Whitewater. He received his M.S. in Industrial Engineering from University of Tehran and his Ph.D. in Management Science and Information Systems from Oklahoma State University. His research interests include health analytics, business analytics, IT privacy, and technology addiction. He is a member of the Association for Information Systems, The Institute for Operations Research and the Management Sciences, and the Decision Sciences Institute. Behrooz has published in journals such as *Journal of the American Medical Informatics Association* (JAMIA), *International Journal of Medical Informatics*, *Expert Systems with Applications*, *Health Informatics Journal*, as well as IS proceedings such as the *Hawaii International Conference on System Sciences* (HICSS) and the *International Conference on Information Systems* (ICIS).

Dr. Dursun Delen is the holder of Spears Endowed Chair in Business Administration, Patterson Family Endowed Chair in Business Analytics, Director of Research for the Center for Health Systems Innovation, and Regents Professor of Management Science and Information Systems in the Spears School of Business at Oklahoma State University (OSU). Dr. Delen has over 30 years of experience in analytics both as a business consultant and university professor. Prior to his academic tenure, he worked for a privately-owned research and consultancy company as a research scientist for five years, during which he led a number of decision support, information systems and advanced analytics related research projects funded by industry and federal agencies including DoD, NASA, NIST and DoE. Dr. Delen has published more than 140 peer-reviewed articles and eight books/textbooks. He is often invited to national and international conferences for keynote addresses, and companies for consultancy engagements on topics related to business analytics, data/text mining, and knowledge management. He regularly serves as chair for tracks and mini-tracks at various business analytics and information systems conferences. Currently, he is serving on more than a dozen journal editorial boards as editor-in-chief, senior editor, associate editor, and editorial board member. He is the recipient of several research and teaching awards including the prestigious Fulbright scholar, regents distinguished teacher and researcher, and Big Data mentor awards.

Amir Hassan Zadeh is an associate professor of MIS at Wright State University. He received his PhD in Management Information Systems from Oklahoma State University, OK, US. His research and teaching interests are in enterprise systems and applications, business analytics, data and text mining, and big data applications. His research works have appeared in journals such as *Decision Support Systems*, *Production Planning & Control*, *Journal of Cases on Information Technology*, and *International Journal of Advanced Manufacturing Technology* among others. He serves as the marketing and communication co-chair of ICIS Decision Support and Analytics (SIGDSA) symposium.