Health Policy and Technology 8 (2019) 7–13

Contents lists available at ScienceDirect

Health Policy and Technology

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

Bringing big data analytics closer to practice: A methodological explanation and demonstration of classification algorithms

Ofir Ben-Assuli[a,*], Tsipi Heart[a], Nir Shlomo[b], Robert Klempfner[c]

[a] Faculty of Business Administration, Ono Academic College, Israel
[b] The Leviev Heart Center, Sheba Medical Center, Ono Academic College, Israel
[c] The Leviev Heart Center, Sheba Medical Center, Israel

ARTICLE INFO

Article history:
Available online 6 January 2019

Key words:
Congestive heart failure
Machine learning
Logistic regression
Boosted decision tree
Support vector machine
Neural network

ABSTRACT

Background: Big data analytics are becoming more prevalent due to the recent availability of health data. Yet in spite of evidence supporting the potential contribution of big data analytics to health policy makers and care providers, these tools are still too complex to be routinely used. Further, access to comprehensive datasets required for more accurate results is complex and costly. Consequently, big data analytics are mostly used by researchers and experts who are far removed from actual clinical practice. Hence, policy makers should allocate resources to encourage studies that clarify and simplify big data analytics so it can be used by non-experts (e.g., clinicians, practitioners and decision-makers who may not have advanced computer skills). It is also important to fund data collection and integration from various health IT, a pre-condition for any big data analytics project.
Objectives: To methodologically clarify the rationale and logic behind several analytics algorithms to help non-expert users employ big data analytics by understanding how to implement relatively easy to use platforms as Azure ML.
Methods: We demonstrate the predictive power of four known algorithms and compare their accuracy in predicting early mortality of Congestive Heart Failure (CHF) patients.
Results: The results of our models outperform those reported in the literature, attesting to the strength of some of the models, and the utility of comprehensive data.
Conclusions: The results support our call to policy makers to allocate resources to establishing comprehensive, integrated health IT systems, and to projects aimed at simplifying ML analytics.

Introduction

Policy makers in the healthcare industry have been allocating substantial resources in the last two decades to Health IT (HIT) implementation [1]. HIT includes Electronic Health Record (EHR) systems, Laboratory Information Management Systems (LIMS), Picture Archiving and Communication systems (PACS), and others [2]. Recently, a new era of HIT has emerged, called big data analytics. Large repositories of historical data can be utilized for advanced analytics, using machine-learning (ML) methods and tools [3]. The results can provide insights and support for decisions related to clinical procedures [4], medication and public health policy [5]. In spite of the emergent value of big data analytics, these tools are still difficult to utilize by health providers or policy makers, and to date are mostly used by researchers [6].

Recently, more user-friendly platforms, such as Microsoft Azure ML, provide tools that do not require high expertise, and can be more widely applied [7]. Nonetheless, users need to understand the methodological rationale (at least to a certain extent) behind these algorithms to facilitate an educated and effective use of these tools.

In this paper, we present key ML techniques and show how they can be used in predicting the early mortality of Congestive Heart Failure (CHF) patients, using data collected from a comprehensive HIT installed at the Sheba Heart Center in Israel. We describe the rationale for each algorithm, compare them in terms of accuracy and other characteristics, and demonstrate the potential value that can be derived from using these techniques. We aim to show that big data analytics can be simpler if more efforts are invested in explaining its principles and methods, and underscore the value of data retrieved from comprehensive HIT. The main contribution of this article is to clarify and simplify big

* Corresponding author.
E-mail addresses: ofir.benassuli@gmail.com (O. Ben-Assuli), tsipi.h@ono.ac.il (T. Heart), nirshlomo81@gmail.com (N. Shlomo), klempfner@gmail.com (R. Klempfner).

https://doi.org/10.1016/j.hlpt.2018.12.003
2211-8837/

data analytics so it can be used by non-experts (e.g., clinicians, practitioners and decision-makers who may not have advanced computer skills). We recommend that policy makers allocate resources to encourage these types of projects.

Congestive Heart Failure (CHF)

Congestive Heart Failure (CHF) is a clinical syndrome characterized by typical signs and symptoms (e.g., breathlessness, swelling of the lower extremities, fatigue, elevated jugular venous pressure and pulmonary edema). The most common etiology of CHF is Ischemic Heart Disease (or Cardiovascular Disease (CVD)) following acute coronary syndrome. Additional etiologies include structural diseases of the heart (e.g., valvular diseases) or abnormalities of the myocardium resulting in reduced cardiac output or elevated intracardiac pressure [8]. Based on data from 2009 to 2012, an estimated 5.7 million Americans 20 years of age or older had CHF. Projections show that the prevalence of CHF will increase 46% from 2012 to 2030, resulting in $>$8 million people $\geq$18 years of age with CHF in the US. About 223 out of 100,000 Americans died of CVD in 2013 [9], and about 20% of all Medicaid CHF patients are readmitted to the hospital within 30 days post first admission [10]. In 2012, the total cost of CHF was estimated at $30.7 billion. Of this total, 68% was attributable to direct medical costs. Projections show that by 2030, the total cost of CHF will increase almost 127% to $69.7 billion from 2012. This is roughly $244 for every US adult. The estimated cost of CHF patients' death is 123.5 billion dollars in the US alone, mostly due to lost productivity [9].

The number of Israelis diagnosed with heart diseases was reported to be 86.1 and 91.4 per 100,000 in 2010, for females and males respectively [11]. An Israeli survey covering 2007–2010 found that 10.2% of all males and 7.1% of all females were diagnosed with some kind of CVD [12]. This is somewhat lower than the US figures, where 12.2% of all males and 10.0% of all females are diagnosed with some kind of CVD [13].

In this study, we used a comprehensive dataset consisting of 7168 patients admitted to the Sheba Medical Center in the last ten years, whose primary diagnosis was CHF. We secured 33 factors about each patient, including extensive demographic details from the Israeli Population Registry, past and present medical history including medication, diagnoses, and laboratory and imagery results. The data were retrieved from six information systems within the hospital (Chameleon, LIMS, PACS, MUSE, MetaVision and Magic, the hospital's legacy systems, as detailed in the next section) as well as the Israeli Population Registry system. The utilization of this dataset was facilitated by the fact that all the sources belong to the same institution, but since the systems are disparate, heavy data manipulation was still required to achieve a unified data representation across all systems, similar to efforts invested in integrating data from different institutions. However, in contrast to retrieving narrow data from several institutions, our sources provided exceptional information depth and breadth about each patient, allowing feature selection from a vast number of factors. These data sources are reliable and thus ensure data accuracy beyond that described in most of the literature.

The Sheba is infrastructure

The data for this study were retrieved from the Sheba Medical Center, the largest Israeli public hospital, located in the center of the country. Sheba is a leading teaching medical center, annually handling over a 1.5 million patient visits and ~200K emergency visits, and conducts more than two million medical tests of all types. Sheba is an intensively computerized hospital, using a comprehensive Chameleon® EMR system that communicates with the Laboratory Information Management System (LIMS) and Picture Archiving and Communication System (PACS) for result notifications and access. The system is used in the ED and in all inpatient departments, and fully replaces all paper-based medical records. This EMR communicates with the administrative ATD (Admit-Transfer-Discharge) system, which is connected to the Israeli Population Registry to facilitate prompt authentication of patients upon presentation at the hospital. In addition, there are specific departmental ISs, such as the Magic system in the Intensive Care Unit.

In recent years, Sheba has set up a research-focused data warehouse, an effort-intensive endeavor involving researchers and dedicated staff. The data warehouse collects data from the following ISs: Chameleon EMR (structured and unstructured elements), LIMS (biochemistry, blood counts, serology, microbiology cultures, biomarkers, genetic tests etc.), data describing imagery results, manual monitored parameters (patient temperature, blood pressure, pulse, saturation, weight, height etc.) keyed into the Magic system, Medical Devices database (ECG, Echocardiography examinations, Cardiac Catheterization, Nuclear Imaging), the National Population Registry, the National Cancer Registry and the Sheba Executive Survey database. These can be classified as clinical, cardiology and administrative IT; i.e., the three categories identified by Bardhan et al. [14].

Method

Background

We used the Azure Machine Learning (ML) platform for the data analysis [15], which implements machine-learning algorithms, data conversion and transformation functions, etc. We employed four known classification models: logistic regression, the Two Class Boosted Decision Tree, the Two Class Support Vector Machine (SVM) and Neural Network (NN), and combined R code to incorporate specific features in each algorithm. The six main steps in our data mining comparison process using AZURE ML were (1) dataset creation, (2) dataset cleaning, (3) dataset column selection for use in the analysis, (4) data splitting into two subsets, one for training and the other for testing. We used 70% of the data for training and 30% for testing according to accepted heuristics (other split values yielded similar results). Identical datasets were used for all the classification models, (5) comparison of the four classification models, (6) testing the prediction accuracy[1] of the trained models.[2] We then evaluated the results of the classification models using R to create the receiver operating characteristic (ROC) curves and C-statistics.[3]

We secured data from 11,360 patients discharged from Sheba between January 2010 and February 2017, with CHF as the primary diagnosis. We excluded patients who died during hospitalization and patients with missing essential clinical data, leaving

[1] **Accuracy**: The proportion of correct predictions vs. the total number of predictions. Accuracy = (True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative)

[2] We chose a **threshold** value of 0.5 for evaluating the accuracy. The threshold is the cutoff value above which a value belongs to one class, and all other values belong to the other class. For example, if the threshold is 0.5, any patient who scored more than or equal to 0.5 is identified as predicted for mortality, and everyone else belongs to non-mortality class.

[3] **ROC**: the receiver operating characteristic curve illustrates the performance of a binary classifier system as its discrimination threshold varies. A ROC plot examines the tradeoff between the two types of positive signals in the models. The horizontal axis represents the ratio of false positives, and the vertical axis shows the ratio of true positives.

C-Statistic: Area under the ROC (also known as AUC) curve (the values lie between 0 and 1) measures the area under the curve plotted with true positives on the y axis and false positives on the x axis. A diagonal line indicates a poor model for predictions; hence, the C-Statistic, for large datasets, should never be lower than 0.5. It is useful as an assessment tool because the single number allows a fair comparison among models of different types for a given dataset (15).

Table 1
Data description—mortality.

Indicator and description	Total	Alive (0)	Death (1)	*p* value
n—Total admissions included	7168	7030	138	
Patient age	75.69 (12.43)	75.58 (12.44)	81.46 (10.74)	<0.001
Gender (female) = 2 (%)	3226 (45.0)	3169 (45.1)	57 (41.3)	0.426
ECHO1_ef; Ejection fraction	47.61 (15.66)	47.59 (15.66)	48.39 (15.75)	0.552
Past history of PCI = 1 (%)	1070 (14.9)	1055 (15.0)	15 (10.9)	0.219
Past history of MI = 1 (%)	1461 (20.4)	1424 (20.3)	37 (26.8)	0.074
Past history of CHF = 1 (%)	3435 (47.9)	3360 (47.8)	75 (54.3)	0.15
Past history of CVA = 1 (%)	935 (13.0)	916 (13.0)	19 (13.8)	0.899
Has a PVD in the background = 1 (%)	475 (6.6)	462 (6.6)	13 (9.4)	0.246
Has Dyslipidemia = 1 (%)	2605 (36.3)	2559 (36.4)	46 (33.3)	0.514
Has hypertension = 1 (%)	3915 (54.6)	3844 (54.7)	71 (51.4)	0.504
Has diabetes = 1 (%)	2381 (33.2)	2338 (33.3)	43 (31.2)	0.669
Has COPD = 1 (%)	849 (11.8)	831 (11.8)	18 (13.0)	0.759
Blood HB level	11.84 (1.95)	11.85 (1.95)	11.31 (1.90)	0.001
Heart function index (0 = low, 1 = preserved)	0.42 (0.49)	0.42 (0.49)	0.41 (0.49)	0.806
Blood pressure in the lung arteries	47.47 (13.00)	47.41 (13.00)	50.38 (12.89)	0.008
GFR- Kidney function indicator	49.67 (30.99)	49.91 (30.97)	37.43 (29.46)	<0.001
Sodium level in the blood	138.42 (5.57)	138.40 (5.50)	139.51 (8.38)	0.02
Hematocrit	37.83 (6.58)	37.85 (6.56)	36.84 (7.31)	0.075
Prior treatment with angiotensin blockers = 1 (%)	5802 (80.9)	5708 (81.2)	94 (68.1)	<0.001
Prior treatment with Aldactone = 1 (%)	3173 (44.3)	3130 (44.5)	43 (31.2)	0.002
First admission length of stay	5.97 (10.72)	5.89 (10.74)	9.96 (8.23)	<0.001
Past Atrial fibrillation = 1 (%)	2094 (29.2)	2044 (29.1)	50 (36.2)	0.083
Diastolic blood pressure	121.86 (27.42)	122.17 (27.32)	106.25 (27.93)	<0.001
Systolic blood pressure	65.30 (15.19)	65.45 (15.10)	58.07 (17.42)	<0.001
Glucose level	132.76 (57.80)	132.74 (57.74)	133.83 (61.07)	0.827
Liver function	39.77 (126.21)	39.69 (127.26)	44.07 (49.21)	0.687
Blood clot function	1.46 (1.00)	1.46 (1.00)	1.51 (1.03)	0.572
Percent change in LV cavity dimensions at the base with systolic contraction	30.15 (85.41)	30.09 (86.23)	33.19 (11.65)	0.673
Weight	78.29 (18.36)	78.37 (18.44)	74.26 (13.70)	0.009
Height	164.58 (12.04)	164.58 (12.09)	164.77 (9.16)	0.853
BMI	28.42 (5.38)	28.44 (5.39)	27.22 (4.62)	0.008
Left ventricular (LV) mass and left ventricular mass indexed to body surface area estimated by LV cavity dimension and wall thickness at end-diastole	113.94 (30.23)	113.99 (30.21)	111.20 (31.07)	0.282

Notes: (1) Data are the mean (±SD) or number of subjects (proportion). (2) The Death column refers to the patients that actually died from all the population. (3) Initials inside the Table: PCI- percutaneous coronary intervention; MI- myocardial Infraction CHF- congestive Heart Failure; CVA- cerebrovascular accident; PVD- peripheral vascular disease; COPD- chronic obstructive pulmonary disease; HB- hemoglobin; GFR- Glomerular filtration rate.

a dataset of 7168 patients. The dataset consisted of the complete CHF-diagnosed population, as opposed to a sample or a cohort filtered by patient characteristics (i.e., age, history, race or income), and contained patient data from various sources installed in this intensively computerized hospital.

Feature selection

The initial database consisted of more than 1200 features. A pre-feature selection process was performed by excluding variables with substantial missing values and textual variables. In addition, highly correlated variables such as consecutive laboratory and diagnostic tests were removed, leaving 200 features available for the feature selection process. The feature selection process was composed of three main steps. The first was clinical importance, according to the senior cardiologists' judgments. The second step was based on a decision tree algorithm using an information gain method to model the importance of the variables. The last step was univariate analysis, using *t* test for continuous variables or a Chi-square test for categorical variables. A *p* value lower than 5% was considered a statistically significant difference between survivor and non-survivor patients. The final dataset consisted of 32 features that were used to perform predictive analytics (Table 1). This feature selection step was performed prior to running the classification algorithm; hence all four algorithms actually used the same dataset with identical features.

We were able to identify patients who died post-discharge, with the exact date of death, by exploiting the unique ID number provided to every Israeli citizen upon birth, and the Sheba ATD system link to the population registry system.

Logistic regression

Our model used a binary-dependent variable, y_i, for each patient *i*, for each variable *x*, representing occurrence of Mortality (1 = Mortality; 0 = non-Mortality). The model is specified as in Eqs. (1) and (2):

$$P(y_i = 1) = \frac{\exp(\beta_o + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon)}{1 + \exp(\beta_o + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon)} \quad (1)$$

Defining $\pi_i = P(y_i = 1)$ and $1 - \pi_i = P(y_i = 0)$, we have

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 (ECHO1\,ef) + \ldots + \varepsilon \quad (2)$$

$$\frac{\pi_i}{1-\pi_i} = \frac{P(y_i = 1)}{P(y_i = 0)} \quad (3)$$

The ratio (3) represents the odds of the event $y_i = 1$ occurring (ε represents the random error).

Two-Class Boosted Decision Tree

The Two-Class Boosted Decision Tree [16,17] is an ensemble learning method in which predictions are based on an entire collection of trees that correct for errors together. It is appropriate for

a binary dependent variable such as predicting mortality. In this method, the second tree modifies the errors of the first tree, the third tree modifies the errors of the first and second trees, and so on. Forecasts are established on the entire set of trees included in the forecast.

A previous study presented superior results with the Boosted Decision Tree compared to other decision trees including Decision forest, the Bagging Decision Tree and the Randomized Decision Tree [18]. The Boosted Decision Tree can improve accuracy at the cost of excluding predicting training cases that are very tough to classify [19,20].

Stiglic et al. [21] conducted comprehensive predictive modeling of the evolution of medical complications using discharge data from pediatric hospitals. They used regularized logistic regression and comorbidity-based features in the first step, and Boosted Decision Trees in the final step of their analysis. They found that there was an improvement in the comprehensibility of the final predictive model but that it was difficult to interpret the models since each Boosted Decision Tree had to be interpreted separately rather than simply merging all the rules obtained during the process of decision tree boosting.

Boosting can be seen as a way of fitting an extended additive model [22].

Two class SVM

Support Vector Machines (SVMs) are supervised learning models used for classification tasks [23,24] and are designed to recognize patterns from large volumes of data. This classifier is useful for predicting binary outcomes that depend on continuous or categorical independent variables.

Given a set of labeled training examples with binary outcome values, the SVM algorithm assigns new examples into one category or the other such that the two categories are divided by the widest gap possible. New examples are then predicted to belong to a category based on which side of the gap they are identified with. There are many successful applications of SVM models ranging from information retrieval to text and image classification.

Our SVM model was trained on the retrospective data for hospital visits. It was calibrated using the parameter optimization tool in AZURE ML. For instance, in the SVM we can calibrate parameters such as number of iterations (type a number that denotes the number of iterations used when building the model) and Lambda (type a value to use as the weight for L1 regularization).

Two-Class Neural Network

A neural network is a set of interconnected layers, in which the inputs lead to outputs by a series of weighted edges and nodes. The weights on the edges are learned when the neural network is trained on the input data. The direction of the graph proceeds from the inputs through the hidden layers, with all nodes of the graph connected by the weighted edges to nodes in the next layer.

Most predictive tasks can be accomplished easily with only one or a few hidden layers. Recent research has shown that deep neural networks (DNN) can be very effective on complex tasks such as image or speech recognition, in which successive layers are used to model increasing levels of semantic depth.

To compute the output of the network for any given input, a value is calculated for each node in the hidden layers and in the output layer. For each node, the value is set by calculating the weighted sum of the values of the nodes in the previous layer and applying an activation function to that weighted sum.

As in previous studies, we used a Two-Class Neural Network (which is a supervised learning method) module to create a model that predicts a target variable that has only two values [25,26].

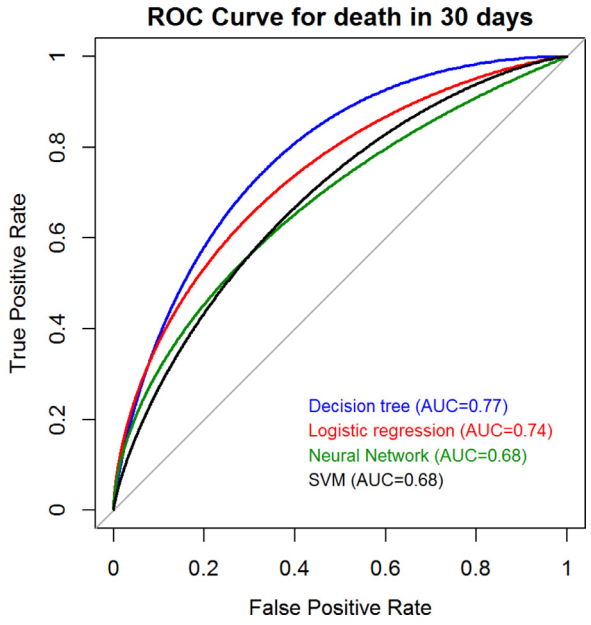

Fig. 1. Mortality within 30 days.

Robustness checksum

Additionally, as a robustness checksum, we verified the mortality within 60 and 90 days post first admission, in addition to mortality within 30 days post first admission.

Results

A feature selection procedure was performed on all factors suggested by clinicians as the most likely to affect early mortality regardless of the classification algorithm used, resulting in 33 factors for each patient (Table 1). These included age, gender, differential diagnoses and current and past health indicators, medications, and lab and imagery results. The average values for twelve indicators differed significantly between the two groups as depicted in Table 1 (significant differences are indicated in bold) for mortality, as this study focused on predicting mortality post first admission.

Fig. 1 displays the ROC curves for the four classification methods predicting mortality, with a C-Statistic ranging from 77.4% to 68.1%. The results (Table 2 and Fig. 1) showed that the best classification model for predicting mortality within 30 days from admission was the Boosted Decision Tree (accuracy = 98.5%; C-statistic = 77.4%), followed by logistic regression (accuracy = 98.4%; C-statistic = 74.3%) which was significantly better [27] than SVM (accuracy = 98.4%; C-statistic = 68.5%) and NN (accuracy = 98.3%; C-statistic = 68.1%).

Additionally, as a robustness checksum, we verified the mortality within 60 and 90 days post first admission and obtained similar findings (Table 2).

In additional prediction of mortality within 60 days from discharge, the C-Statistic ranged from 64% to 74.6%. The results showed that the best classification model for predicting mortality within 60 days from admission was logistic regression (accuracy = 94.8%; C-statistic = 74.6%), followed by SVM (accuracy = 94.9%; C-statistic = 67.9%) and the Boosted Decision Tree (accuracy = 94.9%; C-statistic = 65.2%) and significantly better than NN (accuracy = 94.4%; C-statistic = 64%).

In the alternative time period of predicting mortality within 90 days from discharge the C-Statistic ranged from 62.3% to 72.1%. The results showed that the best classification model for predicting mortality within 90 days from admission was logistic

Table 2
Prediction Mortality for various time periods: comparison across all models (on the Testing Dataset).

Model	Boosted Decision Tree		Logistic regression		Support Vector Machine (SVM)		Neural Network (NN)	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Within 30 days	98.5	77.4	98.4	74.3	98.4	68.5	98.3	68.2
Within 60 days	94.9	65.2	94.8	74.6	94.9	67.9	94.4	64
Within 90 days	92.6	62.3	92.7	72.1	92.6	63.3	91.1	62.8

regression (accuracy = 92.7%; C-statistic = 72.1%), followed by SVM (accuracy = 92.6%; C-statistic = 63.3%) and the Boosted Decision Tree (accuracy = 92.6%; C-statistic = 62.3%) and NN (accuracy = 91.1%; C-statistic = 62.8%).

Discussion

These findings suggest that the objectives set out in this research were partially achieved. We showed that comprehensive and extensively used health IT in hospital settings provide a substantial number of parameters and a potential for prediction improvement and discovery of additional affecting factors (see Table 1).

Several tools were previously introduced to assess risk of mortality and adverse outcomes in CHF patients. The ADHERE Classification and Regression Tree (CART) Model incorporates three routinely measured variables obtained upon hospital admission (SBP, blood urea nitrogen, and serum creatinine). The model stratified risk for in-hospital mortality in Acutely Decompensated Heart Failure (ADHF) patients, with an accuracy of 75.7% in the validation cohort [28]. Another frequently used risk score is the Enhanced Feedback for Effective Cardiac Treatment (EFFECT). This score is designed to predict the propensity for mortality 30 days and at one year in elderly patients with CHF. The EFFECT HF model for 30-day mortality had a C-Statistic of 69%, and the BI-EFFECT index yielded a C-Statistic of 75% [29]. The best results here showed a C-Statistic of ~77%) for early mortality, outperforming current methods, in spite of the relatively small number of positive cases (i.e. early death) in our dataset.

There has recently been a rapid increase in the number of research publications designed to construct prediction models for risk of mortality or readmission in the early post discharge period. Most of the strongest predictors of mortality risk have been reported to be renal function variables (creatinine or GFR), BNP, history of heart failure, age and hemodynamics (blood pressure, pulse) and echocardiographic data (ejection fraction, systolic pulmonary arterial pressure, etc.) [30]. Our results are consistent with previous studies, and underscore the importance of "objectives" variables such as laboratory and diagnostic findings in predicting 30-day post discharge mortality [31,32]. These results emphasize the advantages of combining administrative and clinical features from multiple sources to create versatile longitudinal data. Previous risk models predicting survival in the CHF were developed from prospective studies inherently based on small sample sizes and available physical and medical histories [33]. For instance, the Seattle Heart Failure Model included 1125 patients with severe heart failure. The model comprised age, gender, systolic blood pressure, NYHA class, weight, concomitant medications, lipid profile and laboratory findings [34].

We successfully compared four machine-learning methods in terms of their predictability power and obtained significance differences [27]. The best classification model for predicting mortality within 30 days from admission was the Boosted Decision Tree, followed by logistic regression, SVM and NN. This fulfills our second objective of comparing and differentiating these methods.

These promising results can be attributed to the prevalence of a vast number of quality factors for each patient, due to the unique HIE and interoperability capacity of the hospital's health IT. We found 12 parameters that significantly differed between patients with early mortality and patients who survived (Table 1). These included medications and blood indicators that were not available in previous studies.

Finally, our main finding in this analysis of 7168 patients admitted with a diagnosis of CHF, is that efforts to store and retrieve vast and accurate medical information can pay off by yielding more precise prediction methods. Although our objectives and methodology are quite similar to several previous studies, most of these did not have the variety and breadth of data we had at our disposal. For example, Bardhan et al.'s [14] seminal work, which used data from 67 hospitals, was based on demographic and admission data. Other studies with similar objectives used mainly or solely clinical data such as laboratory and imagery tests [35–37]. In this study, we were able to choose the key parameters from a vast set of demographic, socio-economic, clinical, medical, pharmaceutical, laboratory and imagery data, and time varying covariates. For example, we found that GFR, a kidney function indicator, sodium level in the blood, previous treatment with Angiotensin blockers and past treatment with Aldacton are factors that significantly differentiated the two groups. These factors are rarely found in the literature and have great potential for both clinical and policy recommendations.

This study focused on the entire patient population from one hospital, with CHF as the primary diagnosis, and thereby addresses the call by Futoma et al. [38] to further investigate machine-learning tools while focusing on a specific group of patients.

Policy implications

In this study, we aimed at encouraging policy makers to realize the importance of bringing big data analytics closer to the point of decision-making. Policy makers can derive two main messages from the findings. First, the more comprehensive and multi-sourced the data, the more accurate the predictions are likely to be [3,39]. This, however, requires extensive investment in HIT at points of care such as emergency departments [40,41], community clinics and hospitals [42]. Significant efforts in this direction have primarily been made in developed countries, and are heavily contingent upon public funding [43]. Policy makers should prioritize further efforts in this direction.

Second, we demonstrated that ML algorithms can be methodologically clarified to encourage their use by non-experts. Clinicians and health decision-makers should make efforts to understand the basics of the methodological rationale behind ML algorithms (at least to some extent), to facilitate the effective use of these tools. This article summarized four of the most frequently used ML methodologies in a relatively easy- to- follow way. Together with the availability of user-friendly platforms as Azure ML, this should constitute a step forward toward enabling its use by practitioners. However, more resources should be allocated to advance the simplification of these tools.

Overall, as ML algorithms increase their power and provide more accurate results, especially in light of the growing prevalence of extensive available databases, policy makers should allocate more resources to facilitate the widespread use of ML

analytics [44]. Moreover, if and when ML platforms and algorithms become more friendly and easy to use for non-experts, policy makers will be able to recommend strengthening evidence-based practice to also include ML analytics. The potential beneficiaries are care providers, patients, as well as the policy makers themselves.

Research limitations and further research

Our dataset was composed of 33 variables about each patient, out of ∼1700 potential items (not all present for each patient). While this is more than reported in most current studies, future studies would benefit from additional variables that could possibly reveal more factors affecting CHF patients' risk classification.

In future steps, additional prediction techniques, such as Random Forests [38] and Naïve Bayes Classifiers could be employed and examined to improve predictive power including True Positive (TP) prediction, using various advanced indicators as goodness of fit measurements. Contingent on data retrieval capabilities, another future research avenue would be to analyze the data as time cohorts (e.g., Bayesian networks and Markov Models) for each patient, rather than cross sectional analysis as done here and in all prior studies. Another limitation of this study is regarding the accuracy levels, especially in the True Negative (TN) side with these specific patients. We would achieve high accuracy also without using a complex model since the vast majority of the patients survive in our data set and only about 2% died within 30 days (138 patients died).

Further steps could include cardiologists' interpretation of the results, especially when more than 33 factors are employed, where they will be asked to comment on the clinical insights provided by the results. Another potential future research avenue would be to replicate previous studies with our algorithms, to compare the prediction power. By previous studies we mean those exploring factors identical to the ones in our dataset, since it is very difficult to obtain the actual datasets used in other studies. If better results are obtained compared to the original studies, it will attest to the strength of our algorithms. If not, this would reinforce the assumption that the improvement stems from data quality.

Acknowledgments

This research was supported by a Grant from the GIF (#I-2499-201.2/2018), the German-Israeli Foundation for Scientific Research and Development. This work has also benefited from discussions at WITS 2017 (27th Annual Workshop on Information Technologies and Systems, Seoul, Korea (December 14–15, 2017)).

Funding

This research was supported by a Grant from the GIF (#I-2499-201.2/2018), the German-Israeli Foundation for Scientific Research and Development.

Competing interests

None.

Ethical approval

4904-18-SMC, IRB Committee, Sheba Medical Center Israel.

References

[1] Wachter RM, Making IT. Work: harnessing the power of health information technology to improve care in England. Report of the nationaladvisory group on health information technology in England. England: the National Advisory Group on Health Information Technology in England; 2016.
[2] Shrestha R. The race for the care: a vision for patient-centered care. Appl Radiol 2015;44(9):34–7.
[3] Wang Y, Kung L, Byrd TA. Big data analytics: uunderstanding its capabilities and potential benefits for healthcare organizations. Technol Forecast Soc Change 2018;126:3–13.
[4] Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff 2014;33(7):1123–31.
[5] Khoury MJ, Ioannidis JPA. Big data meets public health. Science 2014;346(6213):1054–5.
[6] Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff 2014;33(7):1163–70.
[7] Elshawi R, Sakr S, Talia D, Trunfio P. Big data systems meet machine learning challenges: towards big data science as a service. Big Data Res 2018;14:1–11.
[8] Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JG, Coats AJ, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur J Heart Fail 2016;18(8):891–975.
[9] Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al. Heart disease and stroke statistics—2016: update. The American Heart Association; 2016.
[10] Wang H, Robinson RD, Johnson C, Zenarosa NR, Jayswal RD, Keithley J, et al. Using the LACE index to predict hospital readmissions in congestive heart failure patients. BMC Cardiovasc Disord 2014;14(1):97–104.
[11] Bugyesky Y. Information and statistics about heart diseases among men and women in israel (in hebrew). Jerusalem: Israeli Knesset; 2014.
[12] Israeli Ministry of Health. Health 2013 (in hebrew). Jerusalem: Ministry of Health; 2014 February 2014. Report No.: Report No. 354.
[13] National Center for Health Statistics. Summary health statistics. Centers for Disease Control and Prevention (CDC); 2015.
[14] Bardhan I, Oh J-h, Zheng Z, Kirksey K. Predictive analytics for readmission of patients with congestive heart failure. Inf Syst Res 2015;26(1):19–39.
[15] Barga R, Fontama V, Tok WH, Cabrera-Cordon L. Predictive analytics with microsoft azure machine learning. Springer; 2015.
[16] Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: practical machine learning tools and techniques. Morgan Kaufmann; 2016.
[17] Freund Y, Schapire RE, editors. A desicion-theoretic generalization of on-line learning and an application to boosting. European conference on computational learning theory. Springer; 1995.
[18] Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach Learn 2000;40(2):139–57.
[19] Freund Y, Mason L, editors. The alternating decision tree learning algorithm. International conference on machine learning; 1999.
[20] Drucker H, Cortes C, editors. Boosting decision trees. Proceedings of the 8th international conference on neural information processing systems. MIT Press; 1995.
[21] Stiglic G, Brzan PP, Fijacko N, Wang F, Delibasic B, Kalousis A, et al. Comprehensible predictive modeling using regularized logistic regression and comorbidity based features. PLoS One 2015;10(12):e0144439.
[22] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). Ann Statist 2000;28(2):337–407.
[23] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.
[24] Teo BK. EXAFS: basic principles and data analysis. Springer Science & Business Media; 2012.
[25] Ou G, Murphey YL. Multi-class pattern classification using neural networks. Pattern Recognit 2007;40(1):4–18.
[26] He H, Wang J, Graco W, Hawkins S. Application of neural networks to detection of medical fraud. Expert Syst Appl 1997;13(4):329–36.
[27] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988:837–45.
[28] Fonarow GC, Adams KF, Abraham WT, Yancy CW, Boscardin W, Committee ASA, et al. Risk stratification for in-hospital mortality in acutely decompensated heart failure: Classification and regression tree analysis. JAMA 2005;293(5):572–80.
[29] Javier Martín-Sánchez F, Gil V, Llorens P, Herrero P, Jacob J, Fernández C, et al. Barthel index—enhanced feedback for effective cardiac treatment (bi—effect) study: contribution of the Barthel index to the heart failure risk scoring system model in elderly adults with acute heart failure in the emergency department. J Am Geriatr Soc 2012;60(3):493–8.
[30] Rahimi K, Bennett D, Conrad N, Williams TM, Basu J, Dwight J, et al. Risk prediction in patients with heart failure: a systematic review and analysis. JACC 2014;2(5):440–6.
[31] Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D, Tu JV. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. JAMA 2003;290(19):2581–7.
[32] Fonarow GC, Adams KF, Abraham WT, Yancy CW, Boscardin WJ, Committee ASA. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. JAMA 2005;293(5):572–80.

[33] Pocock SJ, Ariti CA, McMurray JJ, Maggioni A, Køber L, Squire IB, et al. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. Eur Heart J 2012;34(19):1404–13.
[34] Kalogeropoulos AP, Georgiopoulou VV, Giamouzis G, Smith AL, Agha SA, Waheed S, et al. Utility of the seattle heart failure model in patients with advanced heart failure. J Am Coll Cardiol 2009;53(4):334–42.
[35] Chin MH, Goldman L. Correlates of early hospital readmission or death in patients with congestive heart failure. Am J Cardiol 1997;79(12):1640–4.
[36] Hernandez AF, Greiner MA, Fonarow GC, Hammill BG, Heidenreich PA, Yancy CW, et al. Relationship between early physician follow-up and 30-day readmission among Medicare beneficiaries hospitalized for heart failure. JAMA 2010;303(17):1716–22.
[37] Krumholz HM, Chen Y-T, Wang Y, Vaccarino V, Radford MJ, Horwitz RI. Predictors of readmission among elderly survivors of admission with heart failure. Am Heart J 2000;139(1):72–7.
[38] Futoma J, Morris J, Lucas J. A comparison of models for predicting early hospital readmissions. J Biomed Inform 2015;56:229–38.
[39] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014;2(3):1–10.
[40] Ben-Assuli O. Electronic health records, adoption, quality of care, legal and privacy issues and their implementation in emergency departments. Health Policy 2015;119(3):287–97.
[41] Ben-Assuli O, Leshno M. Using electronic medical records in admission decisions: a cost effectiveness analysis. Decis Sci 2013;44(3):463–81.
[42] Sheikh A, Sood HS, Bates DW. Leveraging health information technology to achieve the “triple aim” of healthcare reform. J Am Med Inform Assoc 2015;22(4):849–56.
[43] Korfkamp D, Gudenkauf S, RohdeEunice M, Kieschke S, Blohm K, Beck A, et al. Opening up data analysis for medical health services: data integration and analysis in cancer registries with CARESS Transactions on large-scale data- and Knowledge-Centered systems XXVI lecture notes in computer Science 9670. H A, K J, W R, B L, M M, editors editors. Berlin, Heidelberg: Springer; 2016.
[44] Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. N Engl J Med 2017;376(26):2507–9.