2019

Real-time Prediction of the Risk of Hospital Readmissions

Nilmini Wickramasinghe

Day Manuet Degano,

Steven McConchie,

32ND BLED ECONFERENCE HUMANIZING TECHNOLOGY FOR A SUSTAINABLE SOCIETY
JUNE 16 – 19, 2019, BLED, SLOVENIA, CONFERENCE PROCEEDINGS
A. Pucihar (ed. et al.)

University of Maribor Press

Real-time Prediction of the Risk of Hospital Readmissions

NILMINI WICKRAMASINGHE, DAY MANUET DELGANO & STEVEN MCCONCHIE

Abstract This study aims to identify predictors for patients likely to be readmitted to a hospital within 28 days of discharge and to develop and validate a prediction model for identifying patients at a high risk of readmission. Numerous attempts have been made to build similar predictive models. However, the majority of existing models suffer from at least one of the following shortcomings: the model is not based on Australian Health Data; the model uses insurance claim data, which would not be available in a real-time clinical setting; the model does not consider socio-demographic determinants of health, which have been demonstrated to be predictive of readmission risk; or the model is limited to a particular medical condition and is thus limited in scope. To address these shortcomings, we built several models to predict all-cause 28-day readmission risk and included Socio-economic Indexes for Areas (SEIFA) data as proxies for socio-demographic determinants of health. Additionally, instead of using insurance claims data, which could require several weeks to process, we built our models using data that is readily available during the inpatient stay or at the time of discharge. The set of default prediction models that were examined include logistic regression, elastic net, random forest and adaptive boosting (Ada Boost). This study examined A not for profit tertiary healthcare organisation from fiscal year 2012-2013 through fiscal year 2017-2018. The out-of-sample results show that all of the models performed similarly and adequately to predict readmission risk.

Keywords: • Risk prediction • Machine learning • Hospital readmission, • length of stay • Real-time prediction •

CORRESPONDENCE ADDRESS: Nilmini Wickramasinghe, PhD, Professor, Swinburne University & Epworth HealthCare, Hawthorn & Richmond, Australia, e-mail: nilmini.work@gmail.com. Day Manuet Degano, Epworth HealthCare, Richmond, Australia, e-mail: Day.Manuet@epworth.org.au. Steven McConchie, Epworth HealthCare, Richmond, Australia, e-mail: Steven.Mcconchie@epworth.org.au.

DOI https://doi.org/10.18690/978-961-286-280-0.6 ISBN 978-961-286-280-0
Dostopno na: http://press.um.si

1 Introduction

Like all OECD countries, Australia is also facing cost pressure regarding delivering high quality care. In the private healthcare sector in Australia unplanned readmissions are 3.1 typically requires the joint analysis of multiple sources of data [2]. However, this can be challenging as data is often incomplete, fragmented and/or consists of misaligned information [3]. This limitation in data quality in turn has hindered epidemiologists to extrapolate demographic information to within plausible limits [4]. Additionally, fragmented data spread across multiple sources makes it difficult for policymakers to compare the relative cost-effectiveness of different interventions [5]. Thus, measuring, gauging and creating benchmarks for unplanned readmission is difficult and yet trying to solve this problem, as is the goal of this research, will have many far reaching consequences.

2 Literature Review and Background

Recent developments in the fields of data warehousing and data science have enabled researchers to contribute to a growing body of knowledge in predictive analytics [3]. In particular, the building, training and application of predictive models to stratify patients into various risk groups based on information from administrative, insurance,
clinical, and government registry sources is becoming a key focus [5]. Such studies are aimed at first aligning complex and sensitive information across multiple sources [6]. This information is then used to identify patients in need of additional healthcare resources by means of various intervention methods [6].

The preponderance of research on predicting unplanned readmissions applies logistic regression models using dichotomous dependent variables [5,8,9,10,11,12,13,14] and occasionally linear regressions [14,11]. Although the variable to be explained is dichotomous, logistic regression can additionally determine the probability of belonging to a certain group, for example, whether a patient is cost intensive (i.e. a likely unplanned readmission or high risk patient) or not (a relatively healthy patient unlikely to have complications) [15]. Compared to logistic regression, the scale level of the dependent variable in linear regression is metric [15]. On the one hand,

the use a dichotomous dependent variable with a well-defined threshold allows for a better comparability. However, the dichotomous dependent variable has the disadvantage that potential cost savings can not directly be assigned [9]. In addition to regression models, classification models such as Support Vector Machine (SVM) and Decision Tree (DT) methods can be applied [16,17,18]. Classification is the assignment of data objects to a suitable class, whereby, for example, the minimization of the classification error or the maximization of the degree of affiliation are used as performance evaluation criteria [19]. In SVMs, data objects are represented as vectors in a ddimensional data space. An SVM looks for a boundary where the objects with different class affiliation are separated as distinctively as possible. This limit is represented by so-called support vectors. In case of more than two attributes, the separating boundary corresponds to a hyperplane [19]. Drosou and Koukouvinos [16] use SVM to find an optimal hyperplane that separates cost-intensive from "regular" patients. However, comparing different classification and predictive models, Moturu, Johnson, and Liu [17] show that SVM have the lowest performance. In their study, Bertsimas et al. [18] utilize DT to classify high-cost patients. The advantage of decision trees lies in the ability to be easily interpreted, where the importance of an attribute is reflected by its proximity to the root node. However, especially for data sets with many attributes, the danger of overfitting occurs [19]. In this case, very large decision trees are created. Although a large decision tree leads to a high classification accuracy on the training data, it does not necessarily lead to a high classification accuracy on the test data [19]. Since the mentioned classification models have not shown a sufficient performance in literature and logistic regression has the advantage of generating probabilities as well, this method is chosen for the predictive analysis. In order to evaluate whether overfitting occurs when learning a classifier, cross-validation of the models is applied.

There are a variety of different influencing factors in literature that increase the likelihood of becoming a costintensive patient. Especially demographic variables are often used as the first factor in predictive analysis, where aspects such as age and gender are known to be reliable predictors [17, 3]. Bertakis and Azari [14] intensively examine the influence of gender in their study and confirm that women are associated with higher costs. Chechulin et al. [3] further verify that good estimates of future costs can be made based on a

person's age. Although pure predictive demographic models perform worse in terms of prognosis quality compared to models with clinical variables, they provide meaningful predictions for the small amount of information available. This allows for categorization at a time when no other information is given [17]. Other important indicators are clinical variables based on the ICD9 and ICD-10 diagnostic codes [3]. Cucciare and O'Donohue [20] further suggest that predictions that include diagnoses show very accurate results. Here, certain chronic diseases, such as diabetes, chronic heart failure (CHF) and chronic obstructive pulmonary disease (COPD), should be studied separately, as these have a major impact on the resulting costs [3]. Hartmann et al. [9] identify accordingly that the metabolic system, especially diabetes, is a trigger for a high number of other diseases and may have long-term effects. Snider et al. [13] support this finding by identifying obesity as an important indicator in their study. This is also related to the body mass index (BMI), sociodemographic variables and other comorbidities. Additionally, people who suffer from a CHF tend to become cost-intensive because they tend to use more healthcare resources of all kinds [21]. Lee et al. [13], define different levels of care, showing that patients with regular care needs are characterized, among other things, by COPD and asthma. In general, diseases can also be summarized in co-morbidity indices and incorporated into the modeling as a predictor [23]. An example is the Charlson Comorbidity Index, which includes diagnoses based on ICD-10 codes [12, 25]. Other relevant predictors include the self-assessment of one's own health status [12, 23], previous healthcare costs [27, 26], resource demands such as number of hospitalizations and number of visits [3, 25], and medication [24, 23]. In the current study we built several models to predict allcause 28-day readmission risk and included Socio-economic Indexes for Areas (SEIFA) data as proxies for sociodemographic determinants of health. Additionally, instead of using insurance claims data, which could require several weeks to process, we focussed on building our own models using data that is readily available during the inpatient stay or at the time of discharge, as the following presents.

3 Methodology

One of the primary objectives of this study is to accurately predict, (ultimately) in real time, the risk of hospital readmission within 28 days of discharge. The following sections describe the underlying data constructions and assumptions that were built into our models.

3.1 Data Preparation

Before developing prediction models, the data set has to be cleaned and prepared. First, variables that have more than 90% missing values or have a constant value over all cases are excluded. Due to input errors in the data set, cases showing inconsistencies across multiple attributes are removed.

3.2 Dataset

The developed models of readmission risk utilised hospital activity, patient characteristics and clinical data, which were derived from six years of admitted patient episode care data, from fiscal year 2012-2013 through fiscal year 2017-2018. These datasets contained episode level information regarding hospital activity, patient characteristics, procedures performed and diagnoses. A separate dataset containing information regarding the specialist, including specialist identification, name and age, was also utilised to develop the model. Eight different SEIFA 2016 scores at a postal code level were incorporated as proxies for socio-demographic determinants of health. This initial dataset contained 202 variables across 926,778 episodes.

4 Outcome Variable

A not for profit tertiary healthcare organisation counts readmissions at the episode level. For the purposes of this study, readmissions were considered for any patient that was readmitted under the following conditions:

- Readmission occurred within 1 to 28 whole days following discharge; and
- readmission occurred for a unique episode; and

- The readmission sequence was discarded.
- Episodes were excluded from the outcome variable for the following reasons:
- Patients were readmitted at a rehabilitation facility; or
- Patients were readmitted at cancelled or hold wards; or
- Patients were readmitted with dialysis or oncology codes; or
- Patients were readmitted with same day mental health treatment; or
- Patients were readmitted with electroconvulsive therapy (ECT) treatment.

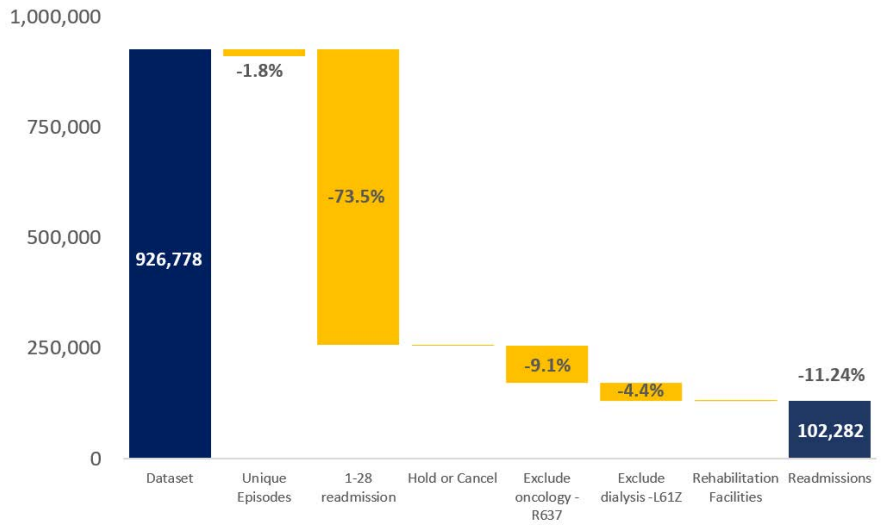

Figure 1. Readmission building block

Of the 926,778 episodes in the initial dataset, 102,282 are identified as readmissions, which represents a readmission rate of 11.24%.

It is important to note that the readmission rate is not included in our model as the dependent (outcome) variable, as a readmission is the final consequence. Because our goal is to predict the risk of readmissions prior to the discharge on the first instance, we instead used the readmission index. The readmission index considers the admission immediately preceding the readmission episode, as quantifying the risks of readmissions prior to

discharge from the initial episode can allow clinicians to identify patients who might benefit from more intensive pre-discharge care.

The readmission variable was calculated according to the formula described in section IV.1.2. Then the index readmission was derived and validated to predict the risk of readmissions within 28 days after discharge from this not for profit tertiary healthcare organisation.

Figure 2. Index – Readmission Concept

4.1 Data Cleaning

To further refine the variables used in the model, we excluded or transformed factors based on the following reasons:

4.2 Unrelated Variables

An extensive consultation process was undertaken with the Clinical Outcomes and Analytics team, the Chair of Health Information Management for a not for profit tertiary healthcare organisation, and external consultants to determine the potential risk factors for readmissions. Empirical evidence suggested that the following variables do not have significant impacts on the risk that a particular patient will be readmitted to the hospital within 28 days of discharge: the division type, the care type, the number of noncertified days of stay, the number of private bed days, the conversion from outpatient to inpatient stays, the conversion from inpatient stays to outpatient stays, the rehabilitation episode type, death after discharge, fund diagnosis related

group (DRG) version, hospital DRG version, principal Commonwealth Medicare Benefits Schedule (CMBS) date and principal CMBS banding.

4.3 Missing feature values

An important number of variables did not include complete records. Careful consideration of methods for dealing with missing data was performed, as failure to appropriately consider missing data can lead to biased results. Variables were generally treated with one of the following methods:

4.4 Elimination

When the missing data represented more than 10% of the total records, the variable was excluded from the modelling dataset. Eliminated variables included the following: unplanned admissions to the ICU, referred by doctor, referred by specialty doctor, referred by doctor at a clinical institute, referred to doctor, referred to specialty doctor, principal shared care doctor clinical institute, miscellaneous code 1, miscellaneous code 2, miscellaneous code 3, miscellaneous code 4, miscellaneous code 5, miscellaneous code 6, miscellaneous code 7, miscellaneous code 8, miscellaneous code 9, miscellaneous code 10, the Australian national subacute and non-acute patient (An-Snap) classification, Snap version, assessment only indicator, date of discharge plan, usual accommodation prior to admission, living arrangement prior to admission, employment status, existing comorbidity, emergency department treating doc 2, emergency department treating doc 3, emergency department treating doc 4, emergency department waiting, emergency department time, triage category, emergency department provisional dx code, emergency department provisional dx, discharge to usual accommodation, policy type and admission patient classification.

4.5 Mean substitution

For continuous variables that contained a low percentage of missing variables, such as the age of the practitioner, the mean value was computed from available cases and was used to replace the missing data values for the remaining cases.

Method of treating missing feature values as special values.

For categorical variables that contained a low percentage of missing variables, such as the insurer group model and SEIFA 2016 factors, the missing variables were treated as new values.

4.6 Inaccuracies

After careful quality inspection of the data, we eliminated the values of discharge age, readmission within 28 days, readmission days, readmission option, height and weight, as these variables were identified as having formulation problems, making their calculations inaccurate.

4.7 Descriptive data

While descriptive data is important for the team to understand the data, these variables were not important for modelling purposes and were therefore excluded: fund DRG description, CMBS description 1, CMBS description 2, CMBS description 3, CMBS description 4, CMBS description 5, CMBS description 6, CMBS description 7, CMBS description 8, CMBS description 9, CMBS description 10, principal diagnosis description, principal coding onset code description, diagnosis coding onset code description 2, diagnosis coding onset code description 3, diagnosis coding onset code description 4, diagnosis coding onset code description 5, diagnosis coding onset code description 6, diagnosis coding onset code description 7, diagnosis coding onset code description 8, diagnosis coding onset code description 9, diagnosis coding onset code description 10, and principal procedure description.

4.8 Insurance claim data

Our primary objective is to develop a model that can be employed in hospital settings to support data-driven discharge interventions to mitigate the risks of hospital readmissions. Thus, we excluded insurance claims data, which could take several weeks to process, as our models requires data that is available during the inpatient stay or at the time of discharge. The variables that fall into this category are the following: fund DRG code, principal

diagnosis code, diagnosis code 2, diagnosis code 3, diagnosis code 4, diagnosis code 5, diagnosis code 6, diagnosis code 7, diagnosis code 8, diagnosis code 9, diagnosis code 10, principal procedure code, procedure code 2, procedure code 3, procedure code 4, procedure code 5, procedure code 6, procedure code 7, procedure code 8, procedure code 9, and procedure code 10.

4.9 Redundant data

The following variables overlap with other relevant factors and were therefore excluded: discharge destination, Local Government Areas (LGA) code, discharge patient classification, ICU hours, discharge doctor clinical institute, reference to doctor clinical institute, principal procedural doctor clinical institute, CMBS code 2, CMBS code 3, CMBS code 4, CMBS code 5, CMBS code 6, CMBS code 7, CMBS code 8, CMBS code 9, and CMBS code 10.

4.10 Feature construction/Transformation

Based on our previous experience, the discovery of meaningful features contributes to a better understanding of the underlying causes of readmissions. Thus, after another extensive consultation process with the Clinical Outcomes and Analytics team, the Chair of Health Information Management for a not for profit tertiary healthcare organisation and external consultants, the following features were derived and/or transformed: admission patient, insurer identifier grouping, marital status, language, age of admitting doctor, age of discharge doctor, age of procedural doctor, age of anaesthetic doctor, indicator of emergency admission, number of emergency procedures, number of procedure codes used, admission month, admission year, discharge year, discharge month, patient age at discharge, number of previous admissions, and number of previous readmissions within 180 days.

4.11 Normalization

As part of our normalization process, we performed discretization on some continuous variables, such as previous readmissions within 180 days. We also attempted to normalize the remaining continuous variables; however, this approach did not improve modelling performance. Therefore, we did not normalize continuous variables in the final dataset.

4.12 De-identification

A crypto-graphical hash function was applied to the following sensitive variables: patient identification, episode identification, insurer group, doctor identification, and patient date of birth. The variables were internally serialized, and we implemented a cyclic redundancy check (CRC) hash function algorithm to compute a compact digest of the serialized object.

5 Patients

To develop a robust risk prediction model, a number of records were removed based on characteristics related to the episode of care. These records were removed to ensure that their inclusion in the modelling dataset did not reduce the robustness of the risk prediction model. These trimmed records generally fell into one of three categories.

The first category included episodes that were considered to be outliers, as their inclusion would disproportionately skew the risk prediction model. These episodes included the following:

- The number of wards for patients that had visited more than four wards;
- The number of anaesthetic doctors for patients with more than three anaesthetic doctors;
- Patients with negative lengths of stay or lengths of stay greater than 41 days for a single episode;
- Patients that spent more than 300 minutes in the operating theatre;
- Patients that visited more than 7 operating theatres for a single episode;
- Patients over 100 years old; and

- Patients that have visited A not for profit tertiary healthcare organisation more than 95 times.

The total number of episodes considered to be outliers represented 5% of the dataset.

The second category included episodes that were removed on the advice of the Clinical Outcomes and Analytics team, as having admission characteristics could not lead to readmission or being generally unrepresentative for the purposes of determining the probability of readmission. This category included the following:

- Episodes related to rehabilitation health admissions in Brighton, Richmond and the Transitional Living Centre.

The final category was related to decisions regarding which episodes were considered out-of-scope or not representative of the patient population. These episodes were trimmed if they included the following characteristics:

- Duplicate episodes; and
- Intersex or indeterminate patients (2 patients in the whole dataset).

6 Modelling

6.1 Feature selection

Feature subset selection is the process of identifying and removing variables that do not have significant impacts on the risk of a particular patient being readmitted to the hospital within 28 days of discharge. We conducted a univariate logistic regression to identify relevant variables.

6.2 Univariate variable selection

This step identified the top-ranked attributes. For categorical variables, the significance of the correlation between each variable and the index readmission was determined using the likelihood ratio test (LRT), using the p values of the fitted logistic regression. In addition, the prevalence, the chi-

squared test and the odds ratio were also considered. For continuous variables, the significants of the correlation between each variable and the readmission outcome index was determined using the LRT, using the p values of the fitted logistic regression. In addition, the odds ratio was considered. For all variables, the response factor was the index readmission, and the explanatory factor was the tested variable. Attributes with significance levels of $p<0.01$ in the univariate analyses were retained for further analyses. In addition, all factors and conditions with prevalence values of less than 1% within the population of patients were excluded from further analyses. The following features were excluded at this stage: ICU days, language v1, language v2, discharge method, admission shift, urgency of admission, discharge month, admission month, discharge day, unplanned theatre visit during episode, admission day, robot use and same-day or overnight stay indicator.

At this stage, the socioeconomic attribute (Decile Index of Relative Socio-economic Advantage and Disadvantage (IRSAD) that most correlated with the index readmission outcome was selected among the following eight variables: Rank IRSAD, Rank Index of Education and Occupation (IEO), Rank Index of Relative Socio-economic Disadvantage (IRSD), Rank Index of Economic Resources (IER), Decile IRSAD, Decile IEO, Decile IRSD and Decile IER, based on the lowest univariate AIC value.

6.3 Correlated variables

Correlation coefficients were obtained among all of the continuous variables A consultation process with the Clinical Outcomes and Analytics team was undertaken to select the most representative variables among heavily correlated variables (<0.30).

- The total number of beds and the total number of wards exhibited a correlation of 0.83. The total number of wards was selected.
- The total number of anaesthetic doctors and the total number of procedure doctors exhibited a correlation of 0.38. The total number of anaesthetic doctors was selected.

- The length of stay and the total number of procedure doctors exhibited a correlation of 0.31. The length of stay was selected.
- The total number of procedures codes and the total number of procedure doctors exhibited a correlation of 0.78. The total number of procedures codes was selected.
- The length of stay and the total number of beds exhibit a correlation of 0.30. The length of stay was selected.
- The admitting doctor age and the discharge doctor age exhibit a correlation of 0.99. The discharge doctor age was selected.

Variables related to the admitting doctor and the discharge doctor were heavily correlated; thus, it was decided that variables related to the discharge doctor should be retained for further analyses.

6.4 Training, testing and validation datasets

A training dataset composed of 80% of the total sample was used to train the models. A validation dataset composed of 20% of the total sample data was used for the unbiased evaluation of suitable models. A testing dataset composed of 10% of the total sample data was used to provide an unbiased evaluation of a final model fit to ensure that the model did not overfit the data.

We ensured that the three datasets followed the same probability distributions among key variables, such as the index readmission.

6.5 Unbalance Dataset

For machine learning problems, differences in prior class probabilities and class imbalances have been reported to hinder the performance of classification algorithms. To account for these potential issues, we tested several resampling techniques, such as under-sampling the majority (normal) class, over-sampling the minority (abnormal) class, random over-sampling examples (ROSE), and synthetic minority oversampling (SMOTE), which have previously been proposed to address class imbalance problems, and compared their effectiveness. The performance of these techniques was

measure by the receiver operating characteristic curve (ROC) method. In previous studies, the results obtained by using similar methods on artificial domains have been linked to the results obtained in real-world domains.

6.6 Classification algorithms

Initially, we experimented with several classic and modern classifiers, including logistic regression, elastic net and random forests. In each case, a 5-fold cross validation was performed.

7 Discussion and Conclusions

This exploratory study served to identify key steps when analysing large healthcare data sets including: defining the index, managing imbalanced data using various techniques and yet achieving a reasonable ROC and assessing various classification algorithms. Crucial insights include the need to focus on index so as to assess ahead of time likelihood of readmission, gender did not play a key role but being alone at home did appear to have an impact.

There were also aspects that might be addressed due to more focussed patient education in some procedures so that bleeding/pain does not automatically mean the need to return to hospital or the emergency department.

While it is exploratory in nature, this study has several contributions to both theory and practice. As noted above we have been able to provide insights into strategies to adopt in order to develop reasonably reliable predictive models using unbalanced data as well as assess the merits of different classification algorithms in the context of data analytics in healthcare. From the perspective of practice, given that today private healthcare organisations in Australia are facing increasing pressures around reducing unplanned readmissions, a necessary first step is to be able to develop robust strategies to best predict likely readmissions at the time of the initial admission and then implement appropriate risk mitigation strategies to avoid the likely unplanned readmissions. Our results have enabled us to progress with this approach for the specific healthcare organisations data and patient population; however, we believe are findings have wider implications and benefits given the move to value-based care in many healthcare systems globally and thereby the need to manage problematic unplanned readmissions in a

systematic and critical fashion. To date, while the need for data analysis, machine learning and deep leaning in the context of healthcare is recognised as important, key findings, algorithms, models and solutions are still not well developed. This study has served to try to assist in this regard. The developed models will now be tested in a large not for profit tertiary healthcare organisation to assess their predictive powers.

Acknowledgements: We acknowledge the assistance from many individuals at Epworth HealthCare as well as Roger Dench and the Team at Fujitsu for their time and contributions to assist us in progressing with this research work. We also are very grateful for the Epworth Research funding received that supported this project. This project was deemed negligible risk with respect to ethical considerations given that de-identified data was used and thus was exempt from ethics.

References

[1] Australian Commission on Safety and Quality in Health Care. The Second Australian Atlas of Healthcare Variation: Hysterectomy hospitalisations 15 years and over [cited 2018 Feb 22]. Available from: URL: https://www.safetyandquality.gov.au/wp-content/uploads/2017/06/3.1-Hysterectomy-1.pdf.

[2] H. D. Sherman, "Hospital efficiency measurement and evaluation. Empirical test of a new technique," Medical care, vol. 22, no. 10, pp. 922–938, 1984.

[3] H. U. Buhl, M. Röglinger, F. Moser, and J. Heidemann, "Big data," Business & Information Systems Engineering, vol. 5, no. 2, pp. 65–69, 2013.

[4] Murray, C.J.L. and Lopez, A.D. (1996), The global burden of disease: A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020 ; summary, Global burden of disease and injury series, v. 1, Published by the Harvard School of Public Health on behalf of the World Health Organization and the World Bank; Distributed by Harvard University Press, [Cambridge, Mass.].

[5] Y. Chechulin, A. Nazerian, S. Rais, and K. Malikov, "Predicting patients with high risk of becoming highcost healthcare users in ontario (canada)," Healthcare policy = Politiques de sante, vol. 9, no. 3, pp. 68–79, 2014.

[6] D. Blumenthal, B. Chernof, T. Fulmer, J. Lumpkin, and J. Selberg, "Caring for high-need, high-cost patients - an urgent priority," The New England journal of medicine, vol. 375, no. 10, pp. 909–911, 2016.

[7] Billings, J., Georghiou, T., Blunt, I. and Bardsley, M. (2013), "Choosing a model to predict hospital admission: an observational study of new variants of predictive models for case finding", BMJ Open, Vol. 3 No. 8, e003352.

[8] J. Hartmann, S. Jacobs, S. Eberhard, T. von Lengerke, and V. Amelung, "Analysing

predictors for future highcost patients using german shi data to identify starting points for prevention," European journal of public health, vol. 26, no. 4, pp. 549–555, 2016.

[9] P. J. Cunningham, "Predicting high-cost privately insured patients based on self-reported health and utilization data," The American journal of managed care, vol. 23, no. 7, pp. e215–e222, 2017. [13] N. S. Lee, N. Whitman, N. Vakharia, G. B. Taksler, and M. B. Rothberg, "High-cost patients: Hotspotters don't explain the half of it," Journal of general internal medicine, vol. 32, no. 1, pp. 28–34, 2017.

[11] J. A. Fleishman and J. W. Cohen, "Using information on clinical conditions to predict high-cost patients," Health services research, vol. 45, no. 2, pp. 532–552, 2010.

[12] L. J. Leininger, B. Saloner, and L. R. Wherry, "Predicting high-cost pediatric patients: derivation and validation of a population-based model," Medical care, vol. 53, no. 8, pp. 729–735, 2015.

[13] B. Li, J. Cairns, J. Fotheringham, and R. Ravanan, "Predicting hospital costs for patients receiving renal replacement therapy to inform an economic evaluation," The European journal of health economics : HEPAC: health economics in prevention and care, vol. 17, no. 6, pp. 659–668, 2016.

[14] S. Rodriguez, D. Munevar, C. Delaney, L. Yang, and A. Tumlinson, "Effective management of high-risk medicare populations," 2014.

[15] J. T. Snider, K. Bognar, D. Globe, D. Ng-Mak, J. Sullivan, N. Summers, and D. Goldman, "Identifying patients at risk for high medical costs and good candidates for obesity intervention," American Journal of Health Promotion, vol. 28, no. 4, pp. 218–227, 2014.

[16] K. D. Bertakis and R. Azari, "Patient gender differences in the prediction of medical expenditures," Journal of women's health (2002), vol. 19, no. 10, pp. 1925–1932, 2010.

[17] J. Behnke, Logistische Regressionsanalyse: Eine Einführung. Methoden der Politikwissenschaft, Wiesbaden: Springer Fachmedien Wiesbaden GmbH, aufl. 2014 ed., 2014.

[18] K. Drosou and C. Koukouvinos, "Proximal support vector machine techniques on medical prediction outcome," Journal of Applied Statistics, vol. 44, no. 3, pp. 533–553, 2016.

[19] S. T. Moturu, W. G. Johnson, and H. Liu, "Predicting future high-cost patients: A real-world risk modeling application," in IEEE International Conference on Bioinformatics and Biomedicine, 2007 X. Hu, ed.), (Los Alamitos, Calif.), pp. 202–208, IEEE Computer Society, 2007.

[20] D. Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G. Wang, "Algorithmic prediction of health-care costs," Operations Research, vol. 56, no. 6, pp. 1382–1392, 2008.

[21] R. M. Müller and H.-J. Lenz, Business Intelligence. eXamen.press, Berlin and Heidelberg: Springer Vieweg, 2013.

[22] M. A. Cucciare and W. O'Donohue, "Predicting future healthcare costs: how well does risk– adjustment work?," Journal of Health Organization and Management, vol. 20, no. 2, pp. 150–162, 2006.

[23] A. J. Rose, "Targeted approaches to improve outcomes for highest-cost patients," Israel journal of health policy research, vol. 6, p. 25, 2017.

[24] A. G. Crawford, J. P. Fuhr, J. Clarke, and B. Hubbs, "Comparative effectiveness of total population versus disease-specific neural network models in predicting

medical costs," Disease management : DM, vol. 8, no. 5, pp. 277–287, 2005.

[25] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.-C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby, and W. A. Ghali, "Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data," Medical care, vol. 43, no. 11, pp. 1130–1139, 2005.

[26] A. Amin, F. Rahim, I. Ali, C. Khan, and S. Anwar, "A comparison of two oversampling techniques (smote vs mtdf) for handling class imbalance problem: A case study of customer churn prediction," vol. 353 of Advances in Intelligent Systems and Computing, pp. 215–225, Cham: Springer International Publishing, 2015.

[27] M. Hofmann and R. Klinkenberg, Rapidminer: Data mining use cases and business analytics applications. Chapman & Hall Crc, 2016.