EPiC Series in Computing

Volume 60, 2019, Pages 94–103

Proceedings of 11th International Conference
on Bioinformatics and Computational Biology

EPiC
Computing

Predicting high-cost patients by Machine Learning: A case study in an Australian private hospital group

Isabella Eigner[1], Freimut Bodendorf[1] and Nilmini Wickramasinghe[2]
[1] Friedrich-Alexander-University Erlangen-Nuremberg, Germany
[2] Swinburne University, Melbourne, Australia
isabella.eigner@fau.de, freimut.bodendorf@fau.de,
nilmini.work@gmail.com

Abstract

Healthcare is considered a data-intensive industry, offering large data volumes that can, for example, be used as the basis for data-driven decisions in hospital resource planning. A significant aspect in that context is the prediction of cost-intensive patients. The presented paper introduces prediction models to identify patients at risk of causing extensive costs to the hospital. Based on a data set from a private Australian hospital group, four logistic regression models designed and evaluated to predict cost-intensive patients. Each model utilizes different feature sets including attributes gradually available throughout a patient episode. The results show that in particular variables reflecting hospital resources have a high influence on the probability to become a cost-intensive patient. The corresponding prediction model that incorporates attributes describing resource utilization achieves a sensitivity of 94.32% and thus enables an effective prediction of cost-intensive patients.

1 Introduction

Resource planning is an important building block for efficient hospital management [1]. In order to support resource planning, large amounts of data can be used, which are gathered daily in the health sector [2]. Applications in this context include the forecast of expenses and the appropriate allocation of budgets. Thus, early measures can be taken to prevent patients from becoming cost-intensive. This way, resource use and costs of these patients can be reduced [3]. Identification and prediction of cost-intensive patients has been an important field of research for several years. On the one hand, the aim is to reduce costs through targeted treatment methods and health care measures [4]. On the other hand, approaches to early prevention can be identified [5] in order to actively promote a good health status in advance and to shorten the patients' history of suffering [3]. Since cost-intensive patients regularly utilize healthcare resources, e.g., through hospital visits or rehabilitation centers, an improvement in the healthcare system primarily requires an improvement in the care of these patients [4]. These improvements can be achieved through optimal coordination and more efficient acute care and aftercare management [6, 7]. The presented paper is based on patient data from a private Australian hospital group. The main goal is to find out which factors contribute to a high predictive power in order to forecast cost-intensive patients. For this purpose, important predictors are identified in literature and constructed from the data set at hand. In order to develop a suitable model for the predictive analysis of this study, the most common models from literature are analyzed and applied to the data set. The paper is structured as follows: For a better understanding of the subject, section two gives an overview of the Australian healthcare system and the definition of high-cost patients. Subsequently, various prediction methods and relevant predictors of existing studies are presented. Section three explains the individual steps of data preparation and the development of four prediction models to identify cost intensive patients based on different feature sets available at various points during the discharge process. Next, the resulting models are evaluated based on their predictive performance. The paper concludes with a critical appraisal, an outlook for future research and a conclusion.

2 Theoretical and conceptual background

2.1 Australian healthcare system

The foundation of the Australian health care system is Medicare, which typically claims 2% of taxable income [8]. Medicare offers free or discounted medical services, lower drug costs, and free government care delivery [9]. Additional needs, however, require private supplementary insurance.

Benefits not funded by Medicare include, for example, costs for ambulance transport, private patient surgery and accommodation, and home care [10]. In order to cover such treatments, supplementary insurance may be provided in the areas of hospital care, general care and ambulance services [10]. Already half of the Australian population has this kind of insurance, with the majority of insured people aged between 60 and 79. This population seeks greater control over their healthcare and choice of benefits and physicians [11]. To claim reimbursements for their services, hospitals need to provide detailed information for each patient episode to the insurer, e.g., Medicare. Based on the clinical and demographic characteristics of a hospital stay, the hospital is reimbursed for an episode as a whole according to predefined diagnosis-related groups (DRG) instead of receiving individual payments per service. Thus, additional treatments in the hospital can lead to excess costs that are not covered by the insurer. Thus, data that is collected for claims purposes can be used to analyse and predict cost-intensive patients to enable timely interventions and reduce unnecessary treatments and costs.

2.2 Cost-intensive patients

Due to the demographic change and the increasingly aging society, the number of cost-intensive patients is increasing massively. Depending on the country and research study, the relationship between high-cost patients and society and their share of total health care costs varies. However, a common statement is that these patients make up 5% of society but account for 50% of total health care costs [4]. Due to this unequal distribution of costs [7], the identification and handling of cost-intensive patients continue to be a priority and represent one of the most important topics in current healthcare systems.

Cost-intensive patients are defined as patients whose costs are in the top 5% to 15% of the total cost distribution [3, 12, 5]. Most frequently, in previous studies, the top decile of the cost distribution is chosen to define cost-intensive patients, which is also used for the present work. Cost-intensive patients often go along with multiple chronic conditions and are treated with various medications [4]. Often, these are older people nearing the end of their lives [13]. In addition, cost-intensive patients often suffer from other diseases, so-called comorbidities, requiring a coordinated treatment, which leads to additional resource strains [7].

2.3 Related work

2.3.1 Methods

Predictive models aim to identify patterns and dependencies in databases in order to predict future events [14]. The majority of research on predicting high-cost patients applies logistic regression models using dichotomous dependent variables [3, 12, 15, 5, 16, 17, 18, 19] and occasionally linear regressions [20, 17]. 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 or not [21]. Compared to logistic regression, the scale level of the dependent variable in linear regression is metric [21]. On the one hand, the use a dichotomous dependent variable with a well-defined threshold allows for a better comparability. On the other hand, the dichotomous dependent variable has the disadvantage that potential cost savings can not directly be assigned [5].

In addition to regression models, classification models such as Support Vector Machine (SVM) and Decision Tree (DT) methods can be applied [22, 23, 24]. 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 [25]. In SVMs, data objects are represented as vectors in a d-dimensional 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 [25]. Drosou and Koukouvinos [22] 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 [23] show that SVM have the lowest performance. In their study, Bertsimas et al. [24] 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 [25]. 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 [25]. 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.

2.3.2 Cost factors

There is a variety of different influencing factors in literature that increase the likelihood of becoming a cost-intensive 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 [23, 3]. Bertakis and Azari [20] 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 [23]. Other important indicators are clinical variables based on the ICD-9 and ICD-10 diagnostic codes [3]. Cucciare and O'Donohue [26] 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. [5] 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. [19] 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 [27]. 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]. Other relevant predictors include the self-assessment of one's own health status [12, 23], previous healthcare costs [28, 26], resource demands such as number of hospitalizations and number of visits [3, 28], and medication [24, 23]. In the following chapter, various prediction models are presented based on the existing data set and their predictive performance is examined.

3 Data analysis

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. This includes, for example, patients that had theatre charges but no operating time or unrealistic values, such as patients with a BMI above 100. In addition, based on results from literature, patients under the age of 18 are not under further consideration [5]. To include the most common and important diseases as individual attributes in the regression model, each selected disease is coded as a dummy variable. On the one hand, the diseases included in the Charlson Comorbidity Index [29] and the ten most common diseases in the present data set are considered, capturing over 60% of all episodes. Taking into account the research objective of identifying the main cost drivers for a broad mass of patients, rare diseases are neglected for further processing. Additionally, comorbidities are specified as such when a patient suffers from at least one of the diseases listed in the Charlson Comorbidity Index. Next, drug groups are assigned for each episode. Furthermore, variables are re-encoded, such as age and BMI. The squared age of the patients is considered as a second variable, since a nonlinear effect of age on the dependent variable is suspected. Thus, you can check whether an additional year of age of an older patient increases the probability of belonging to the group of cost-intensive patients. For the BMI, the individual values of the patients are assigned to the six categories of underweight, normal and overweight as well as obesity grade 1 to 3, thus forming a categorical variable [30]. Since the data set is made up of individual hospitalizations, data for multiple episodes are available for some patients with several hospital stays during the recording period. As there is no information whether a patient was treated in a different hospital between the individual stays, each episode is considered individually instead of aggregating the information for each patient. Based on the total cost of all present episodes, the top decile is used to represent the group of cost-intensive patients. Dividing the dataset into 10% cost intensive and 90% non-cost-intensive patients leads to an uneven distribution of the classes, which can lead to a reduction in the performance of logistic regression [31]. To counter this issue, random undersampling is performed to achieve an even distribution of cost-intensive and non-expensive patients [32]. This means that the number of cases from the majority class (non cost-intensive patients) are reduced to approach the number of cases in the minority class (cost-intensive patients). Finally, highly correlated variables are excluded for further consideration. After data cleaning and preparation the dataset comprises 195,032 episodes.

3.2 Model development

Based on the available data set, four logistic regression models are considered, utilizing different attribute sets according to their subsequent availability during a hospital stay. Model 1 ("socio-demographics") contains only the socio-demographic variables gender, age, squared age and the BMI of the respective patient. This information is easy to collect and is already available when the patient is admitted to the hospital. The second model ("diagnoses") contains the socio-demographic features as well as variables concerning the diagnosis and course of the disease. These variables are likely to be measured at a later time during the hospital stay, e.g., after the first assessment and more complex to collect. Besides these initial features, model 3 ("resource utilization") also includes attributes on the patient's resource utilization during hospitalization. These include, among others, the length of stay, single room occupancy or the duration of all operations. This information is rather collected towards the end of the hospital stay and thus requires more effort and less time for intervention. In the fourth and last model ("total"), all variables from the previous prediction models are combined. Variables that assess a patient's state of health as well as previous healthcare costs are not present in the existing dataset and can therefore not be validated. In order to prevent overfitting, cross-validation is applied for each model. To evaluate the predictive performance of each model, both the C-statistic (measure of goodness of fit) as well as the sensitivity (proportion of correctly identified cost-intensive patients) is used.

3.2.1 Model 1 (socio-demographics)

Model 1 achieves a C-statistic of 0.643 and a sensitivity of 55.36%. This means that 55.36% of the patients who are in fact cost-intensive are also identified by the model. Furthermore, all recorded variables have highly significant effects (p <0.000). However, the signs of the coefficients and thus the direction of the influence differs in two cases from the predicted direction as described in the literature. Thus, in the present data set, both the age (-0.040, p = 0.000) and the dummy variable for women (-0.224, p = 0.000) show a negative sign and thus have a negative influence on the probability of becoming a cost-intensive patient. The other variables BMI classification (0.416, p = 0.000) and ageˆ2 (0.001, p = 0.000) each show a positive significant effect. This means that people assigned to a higher BMI class are more likely to be cost-intensive patients and the influence of age increases exponentially.

3.2.2 Model 2 (diagnoses)

In addition to the socio-demographic information, the influence of diagnoses and medication on the prediction of cost-intensive patients is investigated in model 2. The C-statistic is considerably higher with 0.894 compared to model 1, with a sensitivity increasing to 79.16% in this model. Overall, a significant effect (p <0.01) is found for 35 of the 37 used variables. The main effects of model 2 are shown in Table 1 below. The first four variables refer to the diagnoses of hip osteoarthritis, knee osteoarthritis, chronic ischemic heart disease and peripheral vascular disease. People suffering from these diseases are significantly more likely to become cost-intensive patients than the compared groups. These diagnoses strongly suggest the use of rehabilitation measures, which is further discussed in model 3. The fifth variable is a dummy variable, which indicates whether a person suffers from any unforeseen side effect from their medications. The presence of such side effects leads to a significantly higher probability to become a cost-intensive patient. According to literature, especially the diagnoses diabetes, CHF and COPD as well as comorbidities in general should be considered explicitly. In this dataset, people diagnosed with diabetes without complications and patients with diabetes with chronic complications are distinguished. The presence of a diabetic disease without complications leads to an increased probability of belonging to cost-intensive group (0.103, p = 0.001), while diabetes with chronic complications presents an even higher significant positive effect (0.428, p = 0.000). CHF (0.650, p = 0.000) and COPD (0.239, p = 0.000) can also be confirmed as significant predictors. Contrary to the literature, comorbidities in general show a significant negative influence (-0.431, p = 0.000).

3.2.3 Model 3 (resource utilization)

In addition to the sociodemographic variables, the third model contains 17 additional attributes that cover different areas of resource utilization. The model shows a C-statistic of 0.972 and a sensitivity of 94.32%, which means that 94.32% of the actual cost-intensive patients are properly predicted by the model. Overall, 17 of the 21 used variables show a significant influence (p < 0.05). The main effects are shown in Table 2.

Attribute	Coefficient	Std. Error	z-Value	p-Value
Coxarthrosis	2.612	0.041	63.071	0
Gonarthrosis	2.279	0.032	70.289	0
Chronic ischaemic heart disease	1.914	0.04	47.633	0
Peripheral vascular disease	1.747	0.056	31.315	0
Medical adverse event	1.306	0.047	27.842	0

Table 1: Model 2 – Attributes

Attribute	Coefficient	Std. Error	z-Value	p-Value
overnight stay	1.722	0.051	33.46	0
length of stay	0.345	0.003	101.5	0
total number of doctors	0.241	0.006	40.25	0
bmi class	0.193	0.017	11.37	0
total number of wards	0.145	0.016	9.326	0

Table 2: Model 3 - Attributes

The first variable is a dummy variable and indicates whether a person was hospitalized, i.e., had at least one overnight stay. This effect shows that people that stay overnight are significantly more likely to become a cost-intensive patients receiving outpatient treatment within a day. The second variable, which indicates the length of stay in the hospital, also shows a highly significant positive effect. The probability of being a high-cost patient increases significantly with each additional day of residence. The third effect shows that the likelihood of becoming a high-cost patient increases significantly as the number of treating doctors increases. Similar effects can be identified for the total number of hospital wards during the patient's stay. Finally, rehabilitation measures can also have a strong impact on a patient's costs. Based on a dummy variable that indicates whether a patient has received rehabilitation measures, the model shows that rehabilitation has a negative influence on becoming a cost-intensive patient (-0.844, p = 0.000).

3.2.4 Model 4 (total)

In model 4 all variables are included in the regression. The C-statistic of the model is the highest of all models with a value of 0.976, achieving a sensitivity of 94.32%. According to this model, 39 of the 54 variables have a significant effect (p <0.05). The five most relevant predictors are shown in Table 3.
The first two and last two variables refer to hip osteoarthritis, osteoarthritis in the knee, peripheral vascular disease, and chronic ischemic heart disease. The presence of any of these diseases increases the likelihood of a patient to become cost-intensive. Another significant effect is the overnight stay, which is included in the resource use category. Thus, the likelihood of being a high-cost patient is increased if the patient spends at least one night in the hospital.

Attribute	Coefficient	Std. Error	z-Value	p-Value
Coxarthrosis	1.851	0.046	40.451	0
Gonarthrosis	1.592	0.038	41.857	0
overnight stay	1.418	0.052	27.194	0
Peripheral vascular disease	1.407	0.069	20.538	0
Chronic ischaemic heart disease	1.298	0.052	25.057	0

Table 3: Model 4 - Attributes

3.3 Model comparison

In order to assess which of the developed models should be preferred for the prediction of cost-intensive patients, this section compares the quality criteria of the individual models (cf. Table 4). First, the classification accuracy is considered, thus, how many percent of the examples in the test set are correctly assigned to the correct class. Starting with an accuracy of 59.85% in model 1, the accuracy increases with each stage, resulting in an accuracy of 93.04% in model 4. Similar results can be seen with precision (proportion of correctly classified positives to the total number of the predicted positives) as well as recall (sensitivity) of each model (cf. Table 4). Model 4 shows the highest values with a precision of 91.98% and a recall of 94.32%. The C-statistic is also an important criterion, that plots the true positive rate (TPR) (sensitivity) against the false positive rate (FPR) (specificity: 1 - FPR) of a model according to different thresholds. Both rates are dependent on one another as a lower threshold will usually lead to a higher number of detected true positives (higher sensitivity), but also a higher number of false positives (lower specificity). The C-statistic ranges from 0.5 to 1, where 0.5 represents a random guess and 1 a perfect prediction. Model 1 has a comparatively low C-statistic of 0.643, whereas model 2 already reaches a value of 0.894. Models 3 and 4 achieve very good scores of 0.972 and 0.976 respectively. Finally, the coefficient of determination R2 is considered, which reflects the proportion of the explained variance by the respective variables. While in model 1 only 6.2% of the variance is explained by the used attributes, model 2 achieves values of 47.8% and models 3 and 4 even explain 76.5% and 78.2% of the variance, respectively. All quality criteria show a consistent trend of an increasing predictive performance over the course of a hospitalization and the resulting increase of available data. Based on these results, sociodemographic data, BMI, and variables on resource utilization are most likely to predict cost-intensive patients.

	Model 1	Model 2	Model 3	Model 4
Accuracy	59.85%	81.88%	92.78%	93.04%
Precision	60.82%	83.71%	91.50%	91.98%
Sensitivity	55.36%	79.16%	94.32%	94.32%
C-statistic	0.643	0.894	0.972	0.976
R^2	0.062	0.478	0.765	0.782

Table 4: Model comparison

4 Limitations and implications

First, it should be noted that only data from the Australian healthcare system was used for this study. Thus, the results cannot be transferred without restriction to other countries. Furthermore, within the Australian healthcare system, the distribution of costs between different hospital groups can vary. In addition, private hospitals included in the study may have a particular impact on the outcomes of predicting high-cost patients, e.g., due to different specializations in the hospital group. Therefore, additional data sets should be used to test the assumptions made in this study to be able to generalize the presented results. In addition, since the analysis excludes persons under the age of 18, the four developed models cannot be used to predict the likelihood of high costs in children and adolescents. The regression covers diseases included in the Charlson Comorbidity Index as well as the 10 most common diseases present in this dataset. Thus, rare diseases are neglected in this study. However, since the main hospital costs are caused by diseases that affect a relatively large number of patients and taking into account the research objective of identifying the main cost drivers, rare diseases are not relevant for this study. Another limitation results from the partial inconsistencies in the data set. Several variables are identified that show input errors and inconsistencies with respect to other related variables. Although these cases are excluded from the analysis, the consistency of the other variables may be questionable. In addition, some variables that are considered relevant predictors in literature, such as the self-assessment of health status, are not present in this dataset and therefore their effect cannot be verified. The impact of these variables could be analyzed by retesting against a more comprehensive dataset. Finally, the data in this study contains only health data from patients regarding their hospital stay. Other relevant information, such as outpatient visits or more detailed data on rehabilitation are not considered. This can also result in high health costs and should therefore be taken into account in the holistic determination of cost-intensive patients. In summary, to make the results more generalizable, data from multiple hospital groups in Australia and other countries should be used for further analysis. In addition, costs that incurred outside the hospital should be considered in order to provide a holistic view of the patient’s healthcare costs.

5 Conclusion

One of the biggest constraints on the predictive power of the models presented is the nature of the cost-intensive patients themselves. In literature as well as in the present work, patients are considered static, meaning that a person, who is considered cost-intensive in one year, continues to be part of the cost-intensive group next year as well. However, this cannot be assumed, as it is a dynamic property of patients. Patient dynamics change and thus should not be overlooked in both patient prediction and treatment. The identification and prediction of cost-intensive patients is of great relevance to the use of early preventive care and resource planning of healthcare facilities. Healthcare is a data-intensive industry that enables the efficient and effective use of predictive models for cost-intensive patients based on large volumes of digitally captured patient data. In particular, previous studies identify clinical variables, self-assessments of one’s health, previous healthcare costs, the use of healthcare resources, and information on administered drugs as efficient predictors of patient costs. In the present study, four logistic regression models are introduced to predict costly patients based on a data set from a private Australian hospital group. The models each consider different groups of variables. Model 3 with the consideration of socio-demographic data, BMI and variables of resource utilization can be regarded as the most efficient regression model. Here, significant resources are an overnight stay in the hospital, the length of stay, as well as the number of attending physicians and hospital wards. The identification of predictors for the affiliation of a person to the group of cost-intensive patients is an important input for efficient hospital management. In the near future, this problem becomes even more important, especially due to the increasingly aging society. In addition, it can be expected that due to the steadily growing amount of data, new and more comprehensive possibilities for the development of predictive models will open up.

References

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

[2] H. U. Buhl, M. Röglinger, F. Moser, and J. Heidemann, "Big data," Business & Information Systems Engineering, vol. 5, no. 2, pp. 65–69, 2013.
[3] Y. Chechulin, A. Nazerian, S. Rais, and K. Malikov, "Predicting patients with high risk of becoming high-cost healthcare users in ontario (canada)," Healthcare policy = Politiques de sante, vol. 9, no. 3, pp. 68–79, 2014.
[4] 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.
[5] J. Hartmann, S. Jacobs, S. Eberhard, T. von Lengerke, and V. Amelung, "Analysing predictors for future high-cost patients using german shi data to identify starting points for prevention," European journal of public health, vol. 26, no. 4, pp. 549–555, 2016.
[6] J. M. McWilliams and A. L. Schwartz, "Focusing on high-cost patients - the key to addressing high costs?," The New England journal of medicine, vol. 376, no. 9, pp. 807–809, 2017.
[7] B. W. Powers and S. K. Chaguturu, "Acos and high-cost patients," The New England journal of medicine, vol. 374, no. 3, pp. 203–205, 2016.
[8] ATO, "Medicare levy," 2017.
[9] Australian Government Department of Human Services, "Medicare services," 2018.
[10] Australian Government Private Health Insurance Ombudsman, "What is covered by medicare?," 2018.
[11] S. Lewis, K. Willis, and M. Franklin, "Explainer: why do Australians have private health insurance?," 2015.
[12] 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: Hot-spotters don't explain the half of it," Journal of general internal medicine, vol. 32, no. 1, pp. 28–34, 2017.
[14] P. Mertens, F. Bodendorf, W. König, M. Schumann, T. Hess, and P. Buxmann, Grundzüge der Wirtschaftsinformatik. Berlin, Heidelberg: Springer Berlin Heidelberg, 2017.
[15] 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.
[16] 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.
[17] 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.
[18] S. Rodriguez, D. Munevar, C. Delaney, L. Yang, and A. Tumlinson, "Effective management of high-risk medicare populations," 2014.
[19] 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.
[20] 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.
[21] J. Behnke, Logistische Regressionsanalyse: Eine Einführung. Methoden der Politikwissenschaft, Wiesbaden: Springer Fachmedien Wiesbaden GmbH, aufl. 2014 ed., 2014.
[22] 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.
[23] 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.
[24] 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.

[25] R. M. Müller and H.-J. Lenz, Business Intelligence. eXamen.press, Berlin and Heidelberg: Springer Vieweg, 2013.
[26] 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.
[27] A. J. Rose, "Targeted approaches to improve outcomes for highest-cost patients," Israel journal of health policy research, vol. 6, p. 25, 2017.
[28] 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.
[29] 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.
[30] WHO, "Body mass index - bmi," 2018.
[31] 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.
[32] M. Hofmann and R. Klinkenberg, Rapidminer: Data mining use cases and business analytics applications. Chapman & Hall Crc, 2016.