–

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

hche | Hamburg Center for Health Economics

Power of machine learning algorithms for predicting dropouts from a German telemonitoring program using standardized claims data

Florian Hofer, Benjamin Birkner, Martin Spindler
Research Paper
Year: 2021
No: 24

Power of machine learning algorithms for predicting dropouts from a German telemonitoring program using standardized claims data

Florian Hofer, Benjamin Birkner*, Martin Spindler

Florian Hofer, Hamburg Center for Health Economics (HCHE), Universität Hamburg, Esplanade 36, 20354 Hamburg, Germany. florian.hofer@uni-hamburg.de, ORCID 0003-2532-0919

***Corresponding author:** Benjamin Birkner, Hamburg Center for Health Economics (HCHE), Universität Hamburg, Esplanade 36, 20354 Hamburg, Germany. benjamin.birkner@uni-hamburg.de, Tel: +49 (0)40-42838-4705; Fax: +49 40 42838-9498, ORCID 0002-4452-4733

Martin Spindler, Hamburg Center for Health Economics (HCHE), Universität Hamburg, Esplanade 36, 20354 Hamburg, Germany. martin.spindler@uni-hamburg.de, ORCID 0002-1294-778

Abstract

Background: Statutory health insurers in Germany offer a variety of disease management, prevention and health promotion programs to their insurees. Identifying patients with a high probability of leaving these programs prematurely helps insurers to offer better support to those at the highest risk of dropping out, potentially reducing costs and improving health outcomes for the most vulnerable.

Objective: To evaluate whether machine learning methods outperform linear regression in predicting dropouts from a telemonitoring program.

Methods: Use of linear regression and machine learning to predict dropouts from a telemonitoring program for patients with COPD by using information derived from claims data only. Different feature sets are used to compare model performance between and within different methods. Repeated 10-fold cross-validation with downsampling followed by grid searches was applied to tune relevant hyperparameters.

Results: Random forest performed best with the highest AUC of 0.60. Applying logistic regression resulted in higher predictive power with regard to the correct classification of dropouts compared to neural networks with a sensitivity of 56%. All machine learning algorithms outperformed linear regression with respect to specificity. Overall predictive performance of all methods was only modest at best.

Conclusion: Using features derived from claims data only, machine learning methods performed similar in comparison to linear regression in predicting dropouts from a telemonitoring program. However, as our data set contained information from only 1,302 individuals, our results may not be generalizable to the broader population.

Highlights

- Identification of patients who might drop out of structured care programs prematurely might help insurers to intervene before such attrition happens
- Machine learning algorithms could offer the opportunity to deal with hidden patterns in highly standardized claims data and thus improve predictions in comparison to simple logistic regression
- Timely availability of claims data and access to medical outcomes from structured care programs might improve future predictions as the current evidence suggests low overall predictive performance to identify potential dropouts

1 Introduction

Chronic diseases have become one of the main drivers of rising health expenditure in developed countries. Many countries have drawn upon the concept of managed care as a measure of cost containment[1]. The most important type of managed care in Germany is disease management programs, which statutory health insurers in Germany are obliged to offer to patients for a variety of chronic diseases [2]. One form that has gained a particular impetus since the introduction of a €300-million-per-year innovation funding program by the German government in 2016 has been telehealth [3]. Although evidence on the effectiveness of several telehealth programs is inconclusive, the results of a considerable number of studies suggest that telehealth can be a valid tool for improving adherence and patient-relevant outcomes [4–11].

However, even the most effective programs may not lead to the desired health outcomes if too many patients drop out of them prematurely. Prediction of individuals who are likely to drop out of a health care program might allow insurers to intervene at an early stage and prevent attrition. Machine learning algorithms might prove beneficial in comparison with standard logistic regression by being able to deal with different data types.

Evidence on the use of machine learning to predict dropouts from structured programs is mostly limited to Massive Open Online Courses [12,13]. Itanie et al. and Liang & Zheng used applicant information gathered before and during participation to predict the likelihood of dropping out. Machine learning has been applied within the health care domain most frequently to improve the precision of estimated survival times [14–22], especially in cancer patients, or to predict the time until certain health-related events occur [23–26]. Only a few studies have applied machine learning techniques to predict the likelihood of individuals participating health care programs, the likelihood of decease at discharge out of intensive care units or their compliance to treatment once they have decided to participate [27–30]. We focus on predicting drop-outs using only routinely available claims data collected one year prior to enrolment of participants.

We aim to investigate whether machine learning algorithms outperform linear regression techniques in predicting the likelihood of an individual dropping out from a telemonitoring program for patients with chronic obstructive pulmonary disease (COPD) using only routinely available claims data. The comparison of different methods will help to assess suitability of routinely collected claims data for prediction of dropouts.

2 Methods

2.1 Telemonitoring intervention

COPD is a chronic, inflammatory lung disease that impairs an individual’s ability to exhale. Progressive development of symptoms and episodes of acute worsening (exacerbations) often requiring hospitalization characterize the disease [31]. The most frequent cause for COPD is tobacco smoking. COPD patients in the telemonitoring program were provided with telemedicine equipment that collected and transmitted information on their vital parameters, such as forced expiratory volume in one second (FEV_1) and blood oxygen saturation, to trained practitioners and nurses at a telemedicine service centre. If vital parameters deviated from a predefined threshold, an alarm was triggered and either patients or their physicians was notified.

2.2 Data

Universal coverage for all mandatory insured persons within the statutory health insurance (SHI) is one of the key features of the German healthcare system [32]. Our analysis was based on patient-level data provided by AOK Bayern, a large statutory health insurer in Germany and consisted of routinely collected claims data. Patient sociodemographics consist of age, gender and insurance contribution class. . Clinical information comprised outpatient and inpatient diagnoses, operations and procedures, drug prescriptions and related costs. Main diagnoses were included as International Statistical Classification of Diseases and Related Health Problems 10 (ICD-10) codes. Operations and procedures were covered by Operation and Procedure Classification System (OPS) codes. Anatomical Therapeutic Chemical (ATC) codes describe prescriptions of pharmaceuticals. The data set also contained information about whether patients were currently participating in disease management or other health programs. Our dataset consists of 1,302 individuals who enrolled in the telemonitoring program between October 2012 and December 2015. Sociodemographics, clinical information and costs one year prior to enrolment are included in our analysis.

2.3 Statistical analysis

The main focus of this study was to investigate whether machine learning algorithms are able to make use of hidden patterns within routinely available claims data. The statistical approach is shown in figure 1. We defined a classification problem where the outcome indicated whether an individual had dropped out of the telemonitoring program until the end of the observation period (positive class = “yes”). We used 80% of the available data as a training set and 20% as a test set.

We compare the performance of machine learning algorithms, namely random forests, gradient boosting machines and neural networks with logistic regression [36]. 10-fold repeated cross validation was applied to avoid overfitting and allow for a first hyperparameter tuning [37]. Further extensive hyperparameter tuning was applied using the feature set that had previously led to the best result for a given method [37]. Downsampling within cross-validation was applied to account for imbalance of classes (36% dropouts) (figure 1).

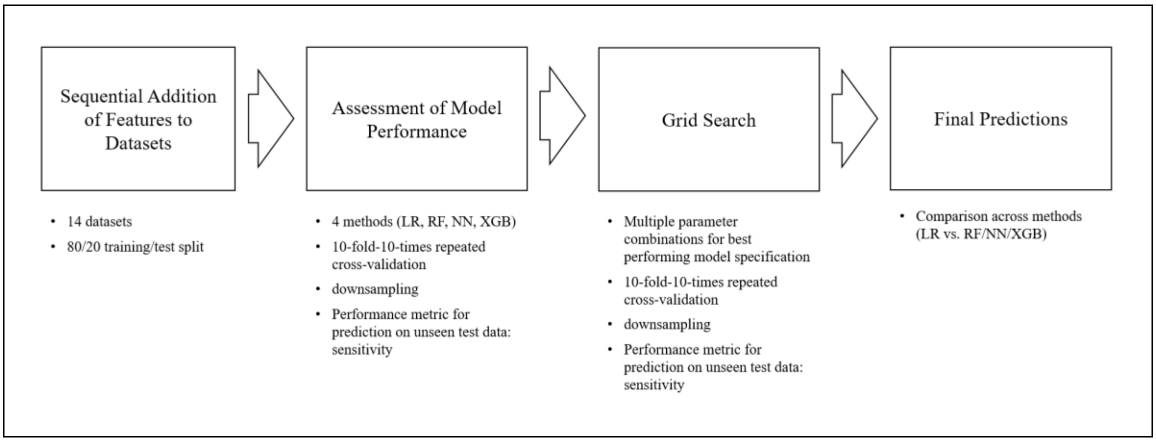

Figure 1 Flow chart of statistical analyses

Our main metric of interest is the correct classification of dropouts, i.e. sensitivity, with the rationale that correct classification of dropouts would be most valuable to the insurer. Subsequently, sensitivity was used as the performance measure of all models for cross-validation, hyperparameter tuning and the final assessment of the predictive power. Additionally, we report the overall performance utilizing the area under the receiver operating characteristics curve (AUC).

2.3.1 Linear regression

Standard logistic regression with all available predictors was used as a baseline out-of-the-box classification model that does not require additional optimization. Logistic regression is expected to result in low variance but high bias when the underlying classification problem is non-linear. Nevertheless, logistic regression should deliver reasonably precise results when the number of predictors is low and features are unrelated [38].

2.3.2 Random forest

The basic idea of decision tree models is to perform discrete splits in the underlying data to form groups that differ as much as possible in terms of the outcome of interest [39]. The algorithm first identifies the predictor with the highest predictive power and splits the data in two groups according to that single predictor. The algorithm then proceeds in the same manner until the addition of a split criterion no longer diminishes the prediction error.

Because the predictive power of a single tree might be poor due to overfitting and high variance, more sophisticated tree-based methods make use of a large amount of single trees [39]. Random forests consist of a certain number of distinct trees is grown, with each tree predicting the outcome. In the case of classification, the final prediction is based on a majority vote of all individual trees. To guarantee the diversity of the trees and prevent overfitting, a random sample of predictors (with replacement) is chosen for each split. Random forests were previously reported as being one of the most accurate predictors in studies that use health-related data [15,18,40]. We chose purity as the split criterion [15,41]. Ten randomly drawn features for each split, a minimum node size of one and 500 trees were used as baseline model parameters. The random forest was implemented using the ranger package within caret for R [42].

2.3.3 Artificial neural networks

Neural networks take features as input parameters and forward them to one or more hidden layers, where the information connected to the input features is weighted to output a prediction [38]. Baseline model parameters were set to the size of five and a decay value of 0.1. We used the neural network package combined with caret in R [43].

2.3.4 Gradient boosting

The intuition of gradient boosting is similar to random forests. An ensemble of weak learning algorithms is combined to create improved prediction modelsGradient boosting is also based on classification and regression trees and is able to use non-linear relationships within the data [44]. However, unlike random forests, gradient boosting machines assign weights to each weak learner instead of a majority vote [19]. Trees within gradient boosting machines are grown sequentially using a random sample of the training data [20]. After each step, the performance of previously grown trees is evaluated and cases that are especially hard to classify are identified. Subsequently a new tree is grown, using higher weights on such cases. This procedure is repeated until the addition of more trees no longer diminishes the prediction error. Baseline model parameters were set to 200 iterations with a maximum depth of 4 per tree. We applied gradient boosting using the XGBoost package in R [45].

2.3.5 Feature selection and hyperparameter tuning

Predictions were based on a base feature set containing information about age, gender and insurance contribution class, the number of main diagnoses falling into the different Elixhauser comorbidity groups [33], the number of drug prescriptions grouped as pharmacy-based metrics (PBM) [34], whether a mental behaviour disorder due to tobacco had been coded and whether the patient had been hospitalized due to COPD in the two years before the start of the program.

We extended the feature set sequentially depending on whether the addition of new data improved the predictive power with regard to the identification of true positives using the previously unseen test dataset. Added features consist of the severity of the underlying disease represented as the individuals' status defined by the Global Initiative for Chronic Obtructive Lung Disease (GOLD) [35] and according to ICD-10 codes, the number of drugs for the treatment of respiratory diseases, participation status in disease management programs two years before enrolment, cost data on inpatient and outpatient treatment and pharmaceuticals and the number of inpatient or outpatient contacts. Lastly, we used a data set containing of the full set of features. If not stated otherwise, all features were based on information gathered from the 12 months before each patient entered the telemonitoring program.

Following the initial prediction process we identified the best feature set for each method and applied hyperparameter tuning based on these datasets. The number of variables at each split was varied for the random forest models from one to 20 in steps of one. Splitting rules were either based on the Gini coefficient or the Hellinger distance measure. Size was varied from one to ten with increments of one unit and decay was varied in 40 steps starting at 0.001 up to two for neural networks. Five different maximum iterations ranging from 50 to 250 and four steps from three to twelve for maximum depth were used for gradient boosting. Additionally, shrinkage was varied from zero to one in ten steps and minimum loss reduction parameters in ten steps from 0.01 to two. Subsample ratios for each tree were set to 33%, 66% and 100%.

3 Results

The participants in the telemonitoring program were 64.4 years old on average and 57.1% were male. In total, 6.8 %, 16.0%, 25.1% and 34.7% had a GOLD status of I, II, III or IV, respectively. For 17.4% of participants, the GOLD status was unknown. Of the 1,302 participants in the telemonitoring program, 472 (36.3%) dropped out. The predictive power of the different machine learning algorithms was highly dependent on the feature set used for prediction.

Overall, none of the machine learning algorithms outperformed simple linear regression with respect to the correct classification of dropouts. However, predictive performance was rather low with a sensitivity of only 56 percent for linear regression and gradient boosting. 52 percent of dropouts was the highest sensitivity achieved by both random forests as well as neural networks.

Interestingly, linear regression shows the highest performance when all available features were used for prediction. Random forest performed best utilizing only the base dataset plus a detailed breakdown of prescribed drugs to treat respiratory diseases with the highest overall AUC of 0.60. Neural networks showed their highest performance when participation in disease management programs was added to the base dataset (AUC: 0.51). Although gradient boosting achieved a sensitivity of 56 percent, the AUC was lowest with 0.51, utilizing the base dataset plus information about inpatient and outpatient contacts per patient. Our results indicate low variance with stable model performance for both the test and the training partition of our datasets (table 1).

The highest differences between performance on the training and test set are found for the simple logistic regression and the Gradient boosting model. Performance dropped for the logistic regression from AUC 0.54 to 0.48 driven by worse identification of true negatives (-5.4 percentage points) without improved identification of true positives. On the contrary, predictive performance increased for random forests with regard to sensitivity decreased a lot by 9.9 percentage points while specificity increased by 2.8 to the achieve the best AUC of 0.60. For more information on the best performing models based on their ability to classify dropouts correctly, see table 1. A graphical comparison of model performance with the receiver operating characteristics curves (ROC) is shown in figure 2.

Table 1 Model performance

Model	Dataset	Training data set (80%)			Test data set (20%)		
		Sensitivity	**Specificity**	**AUC**	**Sensitivity**	**Specificity**	**AUC**
Logistic Regression	All available data	0.554	0.526	0.540	0.560	0.472	0.484
Random Forest	Base feature set + more detailed information about PBM Grooup 21	0.619	0.542	0.588	0.520	0.570	0.597
Artificial Neural Network	Base feature set + information of DMP participation	0.499	0.543	0.539	0.520	0.540	0.512
Gradient Boosting	Base feature set + inpatient/outpatient contacts	0.575	0.526	0.561	0.560	0.549	0.509

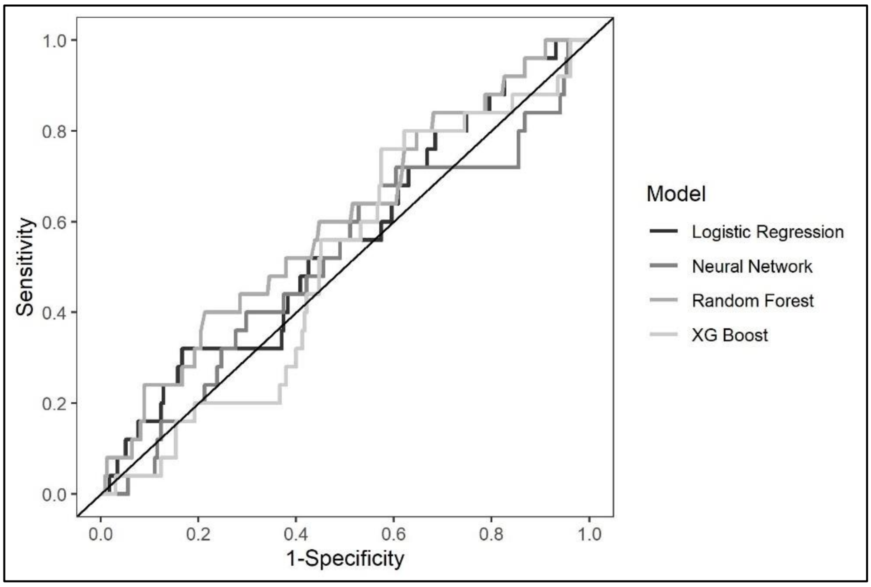

Figure 2 Receiver Operating Characteristics Curves

3.1 Sensitivity Analysis

We investigated the predictive performance of all models in relation to the classification of dropouts with respect to time by defining dropouts according to less than 90, 180 or 270 days (table 2). Sensitivity was highest for a time-horizon of maximum 180 days until dropping out with 68.3% for random forests. The overall predictive performance with the baseline model specifications was best for a maximum of 180 days and decreased after 270 days again. While the results of logistic regression and gradient boosting were largely stable across different time horizons, machine learning algorithms showed higher variability. Neural networks and random forests appear to be able to improve on their predictive power for shorter time horizons (90 days: 56.0%/60.0%, 180 days: 61.0%/68.3%).

Table 2 Sensitivity Analyses with respect to time until dropout

Model	**Dataset**	**Up to 90 days until dropout**				**Up to 180 days until dropout**				**Up to 270 days until dropout**			
		Training (80%)		**Test (20%)**		**Training (80%)**		**Test (20%)**		**Training (80%)**		**Test (20%)**	
		Sensi-tivity	**Speci-ficity**	**Sensi-tivity**	**Speci-ficity**	**Sensi-tivity**	**Speci-ficity**	**Sensi-tivity**	**Speci-ficity**	**Sensi-tivity**	**Speci-ficity**	**Sensi-tivity**	**Speci-ficity**
Logistic Regression	All available data	0.554	0.526	0.560	0.472	0.561	0.547	0.537	0.607	0.540	0.546	0.538	0.548
Random Forest	Base feature set + more detailed information about PBM Grooup 21	0.543	0.557	0.600	0.562	0.540	0.535	0.683	0.598	0.518	0.540	0.635	0.534
Artificial Neural Network	Base feature set + information of DMP participation	0.516	0.527	0.560	0.557	0.511	0.527	0.610	0.584	0.525	0.520	0.365	0.601
Gradient Boosting	Base feature set + inpatient/outpatient contacts	0.562	0.533	0.440	0.494	0.532	0.526	0.537	0.644	0.511	0.524	0.558	0.572

4 Discussion

In this study, we compared the predictive power of linear regression approaches to more advanced machine learning algorithms. When comparing only AUC results, machine learning algorithms show increased performance in comparison to linear regression. However, as the main focus of this study was to predict whether an individual would leave the telemonitoring program prematurely, we focused on the number of correctly classified dropouts as the main performance measure. In this regard, linear regression and Gradient boosting had the highest predictive power and were able to classify approximately 7% more dropouts correctly than was linear regression. The highest sensitivity, of 56.0%, was achieved by linear regression and Gradient boosting. All machine learning algorithms performed better than regressions models in classifying individuals who did not discontinue the program resulting in overall higher AUC. Summarizing, the predictive performance measured in terms of sensitivity was mediocre at best.

One might argue that a focus on sensitivity for model selection could easily be solved by a model which always predicts the positive class and thus identifies all true positives correctly. We chose to focus on sensitivity as the main outcome of interest while still regarding the trade-off between both outcomes. We argue that knowledge about potential dropouts might be the most useful parameter from the SHI perspective. If all participants of a program have to be contacted, the intervention will not be useful for patients who would have not been at risk of dropping out. Being able to identify subpopulations not solely based on the enrolment status to a program or on the main diagnoses offers the opportunity to insurers of targeted interventions to facilitate adherence to structured care programs. This way patients who are classified of being at risk of dropping out might be encouraged to continue with a treatment while adherent patients will not be bothered by needless contacts.

Previous studies have reported that random forests are among the best performing algorithms in predicting survival times using clinical data [15,18,30]. We find that although random forest show the overall best performance, the predictive power is limited. One possible explanation is that our data set was significantly smaller than those used in other studies that examined health related data. Fernandes-Delago et al. found that while random forests were among the best performers for a large amount of data sets, they also needed a large number of data points to perform adequately [40].

As has been observed previously by other researchers, the machine learning algorithms applied in this study would potentially perform better when using data sets containing more features [20]. Because only 1,302 individuals participated in the COPD telemonitoring program, the number of features included should also be rather small to avoid overfitting. We dealt with potential overfitting by repeated 10-fold cross-validation and by application of methods which are known to be less prone to overfitting due to automated variable selection (i.e. Random Forest, Neural Network, Gradient boosting).

We applied random downsampling of the majority class within cross-validation to improve the predictive performance with regard to the minority class. Application of resampling methods within cross-validation is recommended when dealing with imbalanced data to avoid overfitting and overly optimistic predictions [37]. Application of downsampling reduces the data available within cross-valdiation folds and might affect the model performance, especially with regard to the size of our dataset. However, the comparison of downsampling, upsampling and synthetic majority oversampling [46] revealed downsampling as the best approach in our case.

Since we only used information that was available before the start of the telemonitoring program, we expected that our predictions would be more accurate for individuals who left the program soon after it started, an observation reported in a study concerned with predicting survival times for cancer patients [20]. No restrictions regarding duration after program enrolment were made for the definition of dropouts for the baseline analyses. We observed increased model performance throughout all model specifications when identifying patients who dropped out within the first 90, 180 and 270 days after enrolment. Best results were achieved when classifying dropouts after 180 days with a sensitivity of 68.3% for random forest models. However, timing of data availability for the highly standardized claims data of German SHI is often heterogeneous in practice.

Previous studies investigating attendance rates or dropouts for COPD programs have done so mostly for programs concerned with rehabilitation or improving self-management [47,48]. Alongside clinical or behavioural aspects (like smoking habits or episodes of depression [2]), they found that individuals' perceptions of their disease or the program they are supposed to attend are important for their willingness to participate. Furthermore, a person's social environment might also influence his or her decision to participate regularly in a care program [47].

This study has several important limitations. First, because the number of individuals enrolled in the telemonitoring program was small, our results are not generalizable. However, when we tried to predict willingness to participate in the program with a data set that was 10 times larger, the performance of machine learning algorithms improved only slightly. The best performing algorithm was a gradient boosting model, which was able to correctly classify 40.43% of those individuals who were willing to participate in the program.

Second, while using claims data for prediction comes close to real life applications for health insurers in Germany, these data were collected originally for reimbursement purposes and might therefore lack important individual features for each participant. For example, Sohanpal et al. found that attitudes towards care programs drive an individual's choices about whether he or she will participate or stay enrolled in a program. Furthermore, the attitudes of an individual's social contacts are also likely to influence such decisions [47].While the telemonitoring centre in our study gathered detailed information at an individual level, some of which might be valuable for predicting dropouts, we were not able to access this data.

Third, we did not apply an exhaustive lists of algorithms but only those that have been most popular for comparable analyses in the past. In addition, simple logistic regression was used as a baseline model without any adjustments to account for non-linear relationships. While this limits the predictive power of the model, it poses as the simplest approach for a classification problem which might help to decide whether additional effort for building a more complex model is desirable or not.

Fourth, applying different weights for correct/false predictions with the help of cost functions might improve the models predictive power with regard to sensitivity and specificity. However, under the assumption of effectiveness of the telemonitoring program, the main objective is to improve the care for patients with COPD by attending the program. Because no final evaluation of the underlying telemonitoring program for COPD patients is available at this point, the choice of whether to penalize false classification should be done by the insurer upon application.

Finally, our dataset did not contain any information on reasons why patients decided to leave the program. Although further insight would enable insurers to even better identify groups of patients with a high risk of leaving the telemonitoring program, these information is not covered by routinely available claims data.

5 Conclusion

Our study is the first to provide insights into the use of machine learning algorithms to predict the dropout behavior of individuals participating in managed-care-type health programs using only routine claims data. More precisely, we were able to predict up to 56.0% of dropouts from a telemonitoring program correctly using highly standardized data that were gathered before the start of the program with linear regression or machine learning algorithms at best. However, all machine learning algorithms were able to classify patients who did not drop out correctly with a specificity of up to 57%. Subsequently, the additional costs of wrongly classifying individuals as dropouts must be weighed against the additional benefits of classifying more dropouts correctly. Additional socioeconomic or medical features and timely availability of data, might improve the predictive power substantially. Although the overall predictive performance was mediocre at best, our analysis shows the need to evaluate and develop future applications of machine learning to highly standardized data sources which are often used in health economics.

References

1 Shmueli A, Stam P, Wasem J, *et al.* Managed care in four managed competition OECD health systems. *Health Policy* 2015;**119**:860–73. doi:10.1016/j.healthpol.2015.02.013

2 Busse R. Disease management programs in Germany’s statutory health insurance system. *Health affairs* 2004;**23**:56–67.

3 Der Innovationsfonds und der Innovationsausschuss beim Gemeinsamen Bundesausschuss – G-BA Innovationsfonds. https://innovationsfonds.g-ba.de/ (accessed 20 May 2019).

4 Pedone C, Lelli D. Systematic review of telemonitoring in COPD: an update. *Pneumonologia i Alergologia Polska* 2015;**83**:476–84. doi:10.5603/PiAP.2015.0077

5 Lee SWH, Chan CKY, Chua SS, *et al.* Comparative effectiveness of telemedicine strategies on type 2 diabetes management: A systematic review and network meta-analysis. *Scientific Reports* 2017;**7**. doi:10.1038/s41598-017-12987-z

6 Yun JE, Park J-E, Park H-Y, *et al.* Comparative Effectiveness of Telemonitoring Versus Usual Care for Heart Failure: A Systematic Review and Meta-analysis. *Journal of Cardiac Failure* Published Online First: September 2017. doi:10.1016/j.cardfail.2017.09.006

7 Klersy C, Boriani G, De Silvestri A, *et al.* Effect of telemonitoring of cardiac implantable electronic devices on healthcare utilization: a meta-analysis of randomized controlled trials in patients with heart failure: Telemonitoring and healthcare utilization. *European Journal of Heart Failure* 2016;**18**:195–204. doi:10.1002/ejhf.470

8 Alharbi NS, Alsubki N, Jones S, *et al.* Impact of Information Technology–Based Interventions for Type 2 Diabetes Mellitus on Glycemic Control: A Systematic Review and Meta-Analysis. *Journal of Medical Internet Research* 2016;**18**:e310. doi:10.2196/jmir.5778

9 Bashi N, Karunanithi M, Fatehi F, *et al.* Remote Monitoring of Patients With Heart Failure: An Overview of Systematic Reviews. *Journal of Medical Internet Research* 2017;**19**:e18. doi:10.2196/jmir.6571

10 Piga M, Cangemi I, Mathieu A, *et al.* Telemedicine for patients with rheumatic diseases: Systematic review and proposal for research agenda. *Seminars in Arthritis and Rheumatism* 2017;**47**:121–8. doi:10.1016/j.semarthrit.2017.03.014

11 Vitacca M, Montini A, Comini L. How will telemedicine change clinical practice in chronic obstructive pulmonary disease? *Therapeutic advances in respiratory disease* 2018;**12**:1753465818754778.

12 Itani A, Brisson L, Garlatti S. Understanding Learner’s Drop-Out in MOOCs. In: Yin H, Camacho D, Novais P, *et al.*, eds. *Intelligent Data Engineering and Automated Learning – IDEAL 2018*. Springer International Publishing 2018. 233–44.

13 Liang J, Li C, Zheng L. Machine learning application in MOOCs: Dropout prediction. In: *2016 11th International Conference on Computer Science & Education (ICCSE)*. Nagoya, Japan: : IEEE 2016. 52–7. doi:10.1109/ICCSE.2016.7581554

14 Panahiazar M, Taslimitehrani V, Pereira N, *et al.* Using EHRs and Machine Learning for Heart Failure Survival Analysis. *Studies in Health Technology and Informatics* 2016;:14.

15 Churpek MM, Yuen TC, Winslow C, *et al.* Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards: *Critical Care Medicine* 2016;**44**:368–74. doi:10.1097/CCM.0000000000001571

16 Bilal E, Dutkowski J, Guinney J, *et al.* Improving Breast Cancer Survival Analysis through Competition-Based Multidimensional Modeling. *PLoS Computational Biology* 2013;**9**:e1003047. doi:10.1371/journal.pcbi.1003047

17 Chen Y, Jia Z, Mercola D, *et al.* A Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index. *Computational and Mathematical Methods in Medicine* 2013;**2013**:1–8. doi:10.1155/2013/873595

18 Ahmad T, Lund LH, Rao P, *et al.* Machine Learning Methods Improve Prognostication, Identify Clinically Distinct Phenotypes, and Detect Heterogeneity in Response to Therapy in a Large Cohort of Heart Failure Patients. *Journal of the American Heart Association* 2018;**7**. doi:10.1161/JAHA.117.008081

19 Obrzut B, Kusy M, Semczuk A, *et al.* Prediction of 5–year overall survival in cervical cancer patients treated with radical hysterectomy using computational intelligence methods. *BMC Cancer* 2017;**17**. doi:10.1186/s12885-017-3806-3

20 Lynch CM, Abdollahi B, Fuqua JD, *et al.* Prediction of lung cancer patient survival via supervised machine learning classification techniques. *International Journal of Medical Informatics* 2017;**108**:1–8. doi:10.1016/j.ijmedinf.2017.09.013

21 Kourou K, Exarchos TP, Exarchos KP, *et al.* Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* 2015;**13**:8–17. doi:10.1016/j.csbj.2014.11.005

22 Tian Y, Li J, Zhou T, *et al.* Spatially varying effects of predictors for the survival prediction of nonmetastatic colorectal Cancer. *BMC Cancer* 2018;**18**. doi:10.1186/s12885-018-4985-2

23 Davoudi A, Ozrazgat-Baslanti T, Ebadi A, *et al.* Delirium Prediction using Machine Learning Models on Predictive Electronic Health Records Data. In: *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*. Washington, DC: : IEEE 2017. 568–73. doi:10.1109/BIBE.2017.00014

24 Mortazavi BJ, Downing NS, Bucholz EM, *et al.* Analysis of Machine Learning Techniques for Heart Failure Readmissions. *Circulation: Cardiovascular Quality and Outcomes* 2016;**9**:629–40. doi:10.1161/CIRCOUTCOMES.116.003039

25 Kurasawa H, Hayashi K, Fujino A, *et al.* Machine-Learning-Based Prediction of a Missed Scheduled Clinical Appointment by Patients With Diabetes. *Journal of Diabetes Science and Technology* 2016;**10**:730–6. doi:10.1177/1932296815614866

26 Chandir S, Siddiqi DA, Hussain OA, *et al.* Using Predictive Analytics to Identify Children at High Risk of Defaulting From a Routine Immunization Program: Feasibility Study. *JMIR Public Health and Surveillance* 2018;**4**:e63. doi:10.2196/publichealth.9681

27 Linden A, Yarnold PR. Estimating causal effects for survival (time-to-event) outcomes by combining classification tree analysis and propensity score weighting. *Journal of Evaluation in Clinical Practice* 2018;**24**:380–7. doi:10.1111/jep.12859

28 Wallert J, Gustafson E, Held C, *et al.* Predicting Adherence to Internet-Delivered Psychotherapy for Symptoms of Depression and Anxiety After Myocardial Infarction: Machine Learning Insights From the U-CARE Heart Randomized Controlled Trial. *Journal of Medical Internet Research* 2018;**20**:e10754. doi:10.2196/10754

29 Rafael-Palou X, Turino C, Steblin A, *et al.* Comparative analysis of predictive methods for early assessment of compliance with continuous positive airway pressure therapy. *BMC Medical Informatics and Decision Making* 2018;**18**. doi:10.1186/s12911-018-0657-z

30 Roumani YF, May JH, Strum DP, *et al.* Classifying highly imbalanced ICU data. *Health Care Manag Sci* 2013;**16**:119–28. doi:10.1007/s10729-012-9216-9

31 Rabe KF, Watz H. Chronic obstructive pulmonary disease. *The Lancet* 2017;**389**:1931–40. doi:10.1016/S0140-6736(17)31222-9

32 Busse R, Blümel M, Knieps F, *et al.* Statutory health insurance in Germany: a health system shaped by 135 years of solidarity, self-governance, and competition. *The Lancet* 2017;**390**:882–97. doi:10.1016/S0140-6736(17)31280-1

33 Elixhauser A, Steiner C, Harris DR, *et al.* Comorbidity measures for use with administrative data. *Medical care* 1998;**36**:8–27.

34 Kuo RN, Dong Y-H, Liu J-P, *et al.* Predicting healthcare utilization using a pharmacy-based metric with the WHO's Anatomic Therapeutic Chemical algorithm. *Medical care* 2011;**49**:1031–1039.

35 Global Strategy for Diagnosis, Management, and Prevention of COPD - 2016 - Global Initiative for Chronic Obstructive Lung Disease - GOLD. http://goldcopd.org/global-strategy-diagnosis-management-prevention-copd-2016/ (accessed 21 Jun 2016).

36 Wolpert DH, Macready WG. No free lunch theorems for optimization. *IEEE Trans Evol Computat* 1997;**1**:67–82. doi:10.1109/4235.585893

37 Santos MS, Soares JP, Abreu PH, *et al.* Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]. *ieee ComputatioNal iNtelligeNCe magaziNe* 2018;**13**:59–76.

38 Kang J, Schwartz R, Flickinger J, *et al.* Machine Learning Approaches for Predicting Radiation Therapy Outcomes: A Clinician's Perspective. *International Journal of Radiation Oncology*Biology*Physics* 2015;**93**:1127–35. doi:10.1016/j.ijrobp.2015.07.2286

39 Lantz B. *Machine learning with R: discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R*. Second edition. Birmingham Mumbai: : Packt Publishing 2015.

40 Fernandez-Delgado M, Cernadas E, Barro S, *et al.* Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? ;:49.

41 Kuhn M, Johnson K. Classification Trees and Rule-Based Models. In: Kuhn M, Johnson K, eds. *Applied Predictive Modeling*. New York, NY: : Springer New York 2013. 369–413. doi:10.1007/978-1-4614-6849-3_14

42 Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J Stat Soft* 2017;**77**. doi:10.18637/jss.v077.i01

43 Günther F, Fritsch S. neuralnet: Training of Neural Networks. *The R Journal* 2010;**2**:30. doi:10.32614/RJ-2010-006

44 Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 2001;**29**:1189–232.

45 Chen T, He T. xgboost: eXtreme Gradient Boosting. ;:4.

46 Chawla NV, Bowyer KW, Hall LO, *et al.* SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 2002;**16**:321–357.

47 Sohanpal R, Steed L, Mars T, *et al.* Understanding patient participation behaviour in studies of COPD support programmes such as pulmonary rehabilitation and self-management: a qualitative synthesis with application of theory. *npj Primary Care Respiratory Medicine* 2015;**25**. doi:10.1038/npjpcrm.2015.54

48 Keating A, Lee A, Holland AE. What prevents people with chronic obstructive pulmonary disease from attending pulmonary rehabilitation? A systematic review. *Chronic respiratory disease* 2011;**8**:89–99.

hche | Hamburg Center for Health Economics

hche Research Paper Series, ISSN 2191-6233 (Print), ISSN 2192-2519 (Internet)

2011/1 Mathias Kifmann and Kerstin Roeder, Premium Subsidies and Social Insurance: Substitutes or Complements? March 2011

2011/2 Oliver Tiemann and Jonas Schreyögg, Changes in Hospital Efficiency after Privatization, June 2011

2011/3 Kathrin Roll, Tom Stargardt and Jonas Schreyögg, Effect of Type of Insurance and Income on Waiting Time for Outpatient Care, July 2011

2012/4 Tom Stargardt, Jonas Schreyögg and Ivan Kondofersky, Measuring the Relationship between Costs and Outcomes: the Example of Acute Myocardial Infarction in German Hospitals, August 2012

2012/5 Vera Hinz, Florian Drevs, Jürgen Wehner, Electronic Word of Mouth about Medical Services, September 2012

2013/6 Mathias Kifmann, Martin Nell, Fairer Systemwettbewerb zwischen gesetzlicher und privater Krankenversicherung, July 2013

2013/7 Mareike Heimeshoff, Jonas Schreyögg, Estimation of a physician practise cost function, August 2013

2014/8 Mathias Kifmann, Luigi Siciliani, Average-cost Pricing and Dynamic Selection Incentives in the Hospital Sector, October 2014

2015/9 Ricarda Milstein, Jonas Schreyögg, A review of pay-for-performance programs in the inpatient sector in OECD countries, December 2015

2016/10 Florian Bleibler, Hans-Helmut König, Cost-effectiveness of intravenous 5 mg zoledronic acid to prevent subsequent clinical fractures in postmenopausal women after hip fracture: a model-based analysis, January 2016

2016/11 Yauheniya Varabyova, Rudolf Blankart, Jonas Schreyögg, Using Nonparametric Conditional Approach to Integrate Quality into Efficiency Analysis: Empirical Evidence from Cardiology Departments, May 2016

2016/12 Christine Blome Ph.D., Prof. Dr. Matthias Augustin, Measuring change in subjective well-being: Methods to quantify recall bias and recalibration response shift, 2016

2016/13 Michael Bahrs, Mathias Schumann, Unlucky to be Young? The Long-Term Effects of School Starting Age on Smoking Behaviour and Health, August 2016

2017/14 Konrad Himmel, Udo Schneider, Ambulatory Care at the End of a Billing Period, March 2017

2017/15 Philipp Bach, Helmut Farbmacher, Martin Spindler, Semiparametric Count Data Modeling with an Application to Health Service Demand, September 2017

2018/16 Michael Kvasnicka, Thomas Siedler, Nicolas R. Ziebarth, The Health Effects of Smoking Bans: Evidence from German Hospitalization Data, June 2018

2019/17 Jakob Everding, Jan Marcus, The Effect of Unemployment on the Smoking Behavior of Couples, May 2019

2019/18 Jakob Everding, Heterogeneous Spillover Effects of Children's Education on Parental Mental Health, July 2019

2019/19 Esra Eren Bayindir, Hospital Ownership Type and Service Provision, a Structural Approach, November 2019

2019/20 Shushanik Margaryan, Low Emission Zones and Population Health, December 2019

2020/21 Barbara Boggiano, Long-ter effects of the Paraguayan War (1864-1870): from male scarcity to intimate partner violence, May 2020

2020/22 Matthias Bäuml, Christian Kümpel, Hospital Responses to the Introduction of Reimbursements by Treatment Intensity in a (Presumably Lump Sum) DRG System, May 2020

2020/23 Philipp Bach, Victor Schernozhukov, Martin Spindler, Insights from optimal pandemic shielding in a multi-group SEIR framework, November 2020

2021/24 Florian Hofer, Benjamin Birkner, Martin Spindler, Power of machine learning algorithms for predicting dropouts from a German telemonitoring program using standardized claims data, June 2021

The Hamburg Center for Health Economics is a joint center of Universität Hamburg and the University Medical Center Hamburg-Eppendorf (UKE).

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

UKE
HAMBURG

hche
Hamburg Center
for Health Economics

Esplanade 36
20354 Hamburg
Germany
Tel: +49 (0) 42838-9515/9516
Fax: +49 (0) 42838-8043
Email: info@hche.de
http://www.hche.de

HCHE Research Papers are indexed in RePEc and SSRN.
Papers can be downloaded free of charge from http://www.hche.de.