Health Care Management Science (2023) 26:626–650
https://doi.org/10.1007/s10729-023-09653-4

Classification of patients with chronic disease by activation level using machine learning methods

Onur Demiray[1] · Evrim D. Gunes[2] · Ercan Kulak[3] · Emrah Dogan[4] · Seyma Gorcin Karaketir[5] · Serap Cifcili[6] · Mehmet Akman[6] · Sibel Sakarya[7]

Received: 10 July 2021 / Accepted: 4 September 2023 / Published online: 12 October 2023

Abstract
Patient Activation Measure (PAM) measures the activation level of patients with chronic conditions and correlates well with patient adherence behavior, health outcomes, and healthcare costs. PAM is increasingly used in practice to identify patients needing more support from the care team. We define PAM levels 1 and 2 as low PAM and investigate the performance of eight machine learning methods (Logistic Regression, Lasso Regression, Ridge Regression, Random Forest, Gradient Boosted Trees, Support Vector Machines, Decision Trees, Neural Networks) to classify patients. Primary data collected from adult patients (n=431) with Diabetes Mellitus (DM) or Hypertension (HT) attending Family Health Centers in Istanbul, Turkey, is used to test the methods. 44.5% of patients in the dataset have a low PAM level. Classification performance with several feature sets was analyzed to understand the relative importance of different types of information and provide insights. The most important features are found as whether the patient performs self-monitoring, smoking and exercise habits, education, and socio-economic status. The best performance was achieved with the Logistic Regression algorithm, with Area Under the Curve (AUC)=0.72 with the best performing feature set. Alternative feature sets with similar prediction performance are also presented. The prediction performance was inferior with an automated feature selection method, supporting the importance of using domain knowledge in machine learning.

Keywords Patient activation · Patient activation measure · Chronic care · Primary care · Machine learning · Binary classification · Logistic regression · Prediction

Highlights

- Patient activation measure is predicted using machine learning algorithms using a novel rich data set with a large feature set. The few existing studies did not consider most of the features we studied.
- We provide insights for health care providers and managers by identifying features that help predict low-activated/high-risk patients.
- We present the prediction with automated feature selection methods and compare it with feature selection using domain knowledge.

✉ Evrim D. Gunes
egunes@ku.edu.tr

Extended author information available on the last page of the article

1 Introduction

Chronic diseases create an increasing burden and are the leading cause of death and disability worldwide. In this era of the chronic disease pandemic [2], there is an imperative for new approaches in healthcare, commonly identified as "patient-centered care" [34]. This requires understanding a patient's needs and self-management skills and providing appropriate care for that person. There is a widespread consensus on the importance of patient engagement for chronic care management [20, 49]. There is mounting evidence showing that self-management and adherence to treatment of the patients help improve health outcomes and reduce costs, as outcomes depend on the daily actions of patients [6, 15, 42].

Therefore, the concept of patient engagement is emerging as a tool that can be used to personalize care processes [19, 27]. It is even claimed to be the "blockbuster drug" of the century as it can help in achieving the "Triple Aim"

(i.e., lower cost, better health outcomes, higher satisfaction) [8]. Patient engagement is crucial for chronic patients' prognosis and improving treatment adherence. Thus, identifying patients at risk of low engagement should be a priority for chronic care management teams to provide these patients with more guidance and support. This becomes more important for resource-constrained settings, where not all patients can be given the same attention due to limited resources.

To this end, we consider low patient engagement as the primary indication of a need for more support in managing their disease (cf. [36, 41]. In this paper, we aim to investigate methods to identify patients with low patient engagement in order to improve chronic care of the patients and increase life expectancy as well as the quality of life.

There are several scales and constructs related to health-related attitude, skills, and behavior of patients, such as Patient Health Engagement [11], health literacy [33], self-efficacy [40], internal locus of control [30], and patient activation. Among these, we focus on the Patient Activation Measure (PAM), which is a scale that represents "patients' motivation, skills, and confidence in managing their own health" [19]. PAM has the advantage of being a broader construct, including both knowledge, skills, and confidence. It also correlates well with health outcomes and healthcare costs; higher activation levels are associated with improved health behaviors, clinical outcomes, and healthcare costs [12, 38]. Such evidence suggests that identifying those patients with lower activation as the more vulnerable group who would benefit from more attention from the physician and the care coordinator and focusing efforts on these patients can be an effective strategy to improve care.

PAM is a commercially licensed tool. PAM survey is a 13-item scale that contains statements about patients' beliefs about self-management responsibility, self-assessed knowledge, and their confidence in managing health-related tasks (see Appendix for the set of questions used in the scale). The result is reported as a score in the range of $0-100$ and an activation level from 1 (lowest) to 4 (highest). The PAM is being used extensively in the US and the UK to support patient engagement, providing evidence for the usefulness of this scale in practice. [17] provides examples of implementations in practice in the US health system. The National Health Service (NHS) in the UK and the Centers for Medicare and Medicaid Services (CMS) in the US have adopted this tool recently in an effort to improve care for chronic patients and help implement alternative payment methods.[1]

While it has been used widely in the US and the UK, practical implementation of the PAM-13 scale can also be challenging for various reasons. Difficulties arise due to time constraints, the difficulty for some adults in the wording of the scale, strict requirements for standard administration of the test [7, 10], and challenges of validating PAM in various languages. When a PAM score is not available, various factors could help characterize a patient and signal the activation level. Thus, a care management team is at risk of being overloaded with information, sometimes giving conflicting signals. Furthermore, targeted interventions to improve disease management at the public health system level can be more cost-effective. To this end, a systematic tool to predict patients with low activation to target efforts both at the clinic and health system level would be useful.

Using machine learning methods for providing decision support to clinicians is a promising path for the future of clinical medicine [31]. There is a growing literature on using machine learning in predicting adherence behavior, focusing on patients with particular conditions (such as heart failure, and diabetes). The literature on patient engagement in general, and Patient Activation Measure in particular, have focused on understanding associations of these constructs with various attributes of people and their environment. Although there is a broad literature on how PAM is related to other measures and its effects on health outcomes, there is limited research on predicting PAM. One notable exception is the work of [36], which illustrates the use of two machine learning algorithms with a limited set of features and limited data (118 patients) to predict PAM level, while their main focus is on measuring the effectiveness of an intervention to improve PAM. We aim to contribute to filling this gap in the literature. This paper seeks answers to two main research questions: Can we effectively use machine learning methods to predict patient activation levels? Which patient features will be most helpful for this prediction? Thus, our objectives in this paper are: (1) to test machine learning methods for improving predictions and (2) to provide practical tools and insights to identify such patients.

One natural question that may arise is whether predicting adherence instead of PAM would be more beneficial. We believe that while interest in adherence is undoubtedly warranted, focusing only on adherence is not enough. Adherence to medical and non-medical treatment reflects only one aspect of patient activation in managing chronic disease. Several dimensions of patient behavior affect health outcomes, such as medication adherence, diet, exercise, or adhering to regular doctor visits. Patient activation is a construct that correlates with these behaviors and health outcomes, and it would be easier to measure this one dimension rather than several adherence behaviors. Patient activation offers the patient more of an active role in controlling their disease, beyond simply "adhering to what is recommended". In addition, focusing on patient activation as opposed to adherence is a more forward-looking approach that may help facilitate preventive actions. For these reasons, we focus on this con-

[1] see the webpage of NHS https://www.england.nhs.uk/2014/05/patient-activation/ and endorsement by the CMS: https://cmit.cms.gov/CMIT_public/ViewMeasure?MeasureId=3277

struct as the primary variable of interest; our first objective is to predict it using machine learning methods.

Our second objective is to identify the patient characteristics that are most useful for the prediction and to understand the relative usefulness of different information sets. Thus we aim to find simple yet powerful information sets that care teams and policymakers can rely on for predicting PAM. This will provide insights for clinics that manage a large group of patients and do not have the time or resources to administer the PAM questionnaire to identify vulnerable patients. We build on the previous body of literature on patient activation and implement machine learning algorithms to develop methods and insights for predicting PAM.

Our third objective is to compare the performance of machine learning algorithms that learn only from data by automated feature selection with approaches that use domain knowledge. Sinha and Zhao [43] compared methods with and without using domain knowledge and concluded that domain knowledge significantly improves classification performance. Wilcox and Hripcsak [50] and [29] are other examples with similar conclusions. We aim to contribute to the emerging discussion of the importance of using domain knowledge in machine learning with our analysis.

We conduct extensive numerical experiments testing alternative feature sets using a toolbox of different algorithms and report our findings. The recent work of [37] is in line with our second objective in spirit. Rao et al. [37] applies machine learning algorithms to predict antibiotic adherence from a large electronic patient record database and comment on the relative usefulness of different features for prediction.

This research is based on primary data collected from Family Health Centers in Istanbul, Turkey, through a survey including several features. This rich dataset will allow us to investigate the importance of the different amounts of information on prediction power. Using this dataset, we apply eight commonly used machine learning algorithms (Logistic Regression, Logistic regression with L1 regularization (i.e., Lasso Regression), Logistic regression with L2 regularization (i.e., Ridge Regression), Random Forest, Gradient Boosted Trees, Support Vector Machines, Decision Trees, and Neural Networks) to classify patients based on PAM level. We then report our results for several combinations of feature sets and discuss the findings. The prediction problem we consider is a binary classification problem, where patients are classified into low or high PAM level. However, since this classification emerges as a result of the PAM level of patients, we also refer to the problem as "predicting PAM levels" or "predicting low PAM" in the remainder of the paper. Our contributions can be summarized as follows:

1. This paper is among the first papers to focus on predicting PAM and illustrate using machine learning algorithms for this purpose.
2. We use a rich, novel data set collected by the research team, which provides an opportunity to understand the possible use of a broader range of features for predicting PAM. The few existing studies did not consider most of the features we studied.
3. We provide insights for health care providers and managers by identifying features that help predict low-activated/high-risk patients.
4. We present the prediction with automated feature selection methods and compare it with feature selection using domain knowledge. This analysis also contributes to the emerging discussion of the importance of domain knowledge in the machine learning literature.

The rest of the paper is organized as follows. Section 2 reviews the literature. Section 3 describes the research methods, including the survey and data collection, and our approach to implementing the machine learning algorithms. In Section 4, we present the results and discuss the findings. Section 5 presents an extension of the analysis for alternative definitions of Low PAM. Section 6 concludes with a summary of the results and discusses limitations and future research directions.

2 Related literature

This research is related to two streams of literature; patient engagement with patient activation measure and prediction of patient engagement with machine learning tools. In this section, we briefly review relevant work in these literature streams.

There is a vast literature on applying predictive analytics in health care and medicine. We name a few examples: [48] uses Support Vector Machines to diagnose breast cancer from mammography images, [4] reviews machine learning applications focusing on the diagnosis, classification, readmissions, and medication adherence in heart failure. There has been an increasing interest in implementing machine learning algorithms to predict patient adherence behavior. Karanasiou et al. [21] uses data from 90 patients to employ several machine learning algorithms to predict the different types of adherence behaviors for heart failure patients. Son et al. [45] applies Support Vector Machines to predict medication adherence in Heart Failure patients, with data from 76 patients, and concludes that this is a promising method based on the best-observed accuracy. A comprehensive review of machine learning methods for heart failure can be found in [46]. Zhou et al. [52] predicts exercise adherence using data from a trial involving 210 patients, designing a method using logistic regression and support vector machines. Rao et al. [37] predicts antibiotics adherence applying machine learning methods. Koesmahargyo et al. [23] uses Extreme

Gradient Boosting to predict medication non-adherence in a large clinical trial (n=4182) and finds the empirically observed adherence as the strongest predictor of future adherence/non-adherence. Wallert et al. [47] uses a Random Forest classifier to predict adherence to internet-delivered psychotherapy, using data from 90 patients. Bonnell et al. [5] applies various machine learning algorithms to predict unhealthy drinking with a large dataset (n=43545) and selects the random forest model to use for prediction based on the area under the receiver operator curve. Rao et al. [37] predicts antibiotic adherence using a random forest classifier with a large database of Electronic Health Records. Queenan et al. [36] predicts PAM level with support vector machines. We use a methodology similar to this stream of literature. We employ alternative machine learning methods and evaluate their performance. In contrast, our target variable will be low PAM level, which has not been given enough attention in this stream of literature.

The literature on patient engagement is extensive; therefore, here, we focus on papers that focus on PAM. Much attention has been given to PAM's association with health outcomes. Greene et al. [13] finds that higher PAM levels are associated with better health outcomes and lower costs, using data from primary care. Hibbard et al. [16] shows evidence that lower activation is associated with higher costs using data from a health system in Minnesota. They also show that the low PAM is a significant predictor of higher costs, even after adjusting for a clinical risk score [17].

There is also some work investigating determinants of the patient activation level. For example, [51] investigates the effect of Primary Health Care related factors such as access, utilization, responsiveness, interpersonal communication, and patient satisfaction. They find that patients who could spend more time with the care provider were more likely to have high activation. Smith et al. [44] studies the association of PAM, Health Literacy, and health outcomes and found that both of these constructs correlate with health outcomes. Lindsay et al. [25] finds that a single PAM level increase is associated with 8.3% lower follow-up costs and concludes that change in PAM can be used as an early signal of changing costs. O'Malley et al. [32] studies PAM in prostate cancer and breast cancer survivors and finds that employment status, household income, ease of access to oncology team and primary care physicians, and perception of time spent with the physician were associated with higher PAM levels. Prothero et al. [35] investigates the gender differences and finds no evidence of the effect of gender on the PAM level. Similar to this stream of literature, our focus is also on PAM Level, and we use logistic regression to observe factors associated with a low PAM level. Unlike the existing work, our focus is on classifying patients using machine learning algorithms to detect those with low PAM.

3 Methods

3.1 Research setting

The data is collected in Istanbul, Turkey, from patients with Hypertension or Type 2 Diabetes Mellitus visiting their Family Health Centers (FHC). Turkey has a publicly funded national health system, where all services in FHCs are free. At the same time, there is no gatekeeping and referral system in the country, and everyone can receive health services in public hospitals with some contribution fee. There is also a private hospital sector, and complementary and general private health insurances are popular. Diagnosis and treatment of chronic diseases are primarily made at the secondary care level by specialist physicians [22]. While there is a vast network of Family Health Centers, patients apply to these clinics primarily for prescription refilling purposes.

3.2 Survey and data collection

The study sample consists of patients with diabetes and/or hypertension, aged 30-75, attending nineteen Family Health Centers (FHC) in the eastern part of Istanbul. Family Health Centers were chosen randomly from two regions, one with a higher socioeconomic level on average than the other. The survey was administered in person by researchers in the FHCs these patients attend. Data was collected from n=431 patients (mean age: 63.6); 65% were women; 67% had more than one chronic condition). The study was approved by the Ethics Committee of Marmara University.

The survey instrument has 38 questions; after removing one redundant question and the features with too many missing values, we are left with 33 features. The features in our data obtained with this survey can be categorized as information related to four different factors; these are labeled (1) Demographics, (2) Habits, (3) Health, and (4) Health Service. A list of features under each category is listed in Table 1. The Appendix shows summary statistics for these features in Table 13.

In addition, the Turkish translation of the PAM scale (translated and validated by [24]) is administered. PAM is scored between 1-4. We label PAM levels 1 and 2 as "Low PAM," and PAM levels 3 and 4 as "High PAM," following the practice of the NHS, which defines levels 1 and 2 as low PAM levels.[2] While another cut-off could be used to define low activation, patients with levels 1 and 2 are considered to be needing more support. Nevertheless, we also provide a sensitivity analysis based on using other cut-off points for classification in Section 5. In our study sample, 44.5% of patients had a low PAM level (PAM level 1 or 2). Figure 1

[2] See https://www.england.nhs.uk/wp-content/uploads/2018/04/patient-activation-measure-quick-guide.pdf

Table 1 Features collected by the survey

Feature Set and Feature Names	Description: Levels of feature (if categorical)
Demographics	
age	
gender	female, male
livesalone	yes, no
***education	no school, primary/secondary school, high school, university and above
income	less than minimum wage, minimum wage, 2 or 3 times the minimum wage or more
occupation	working outside home, not working outside home, retired
private-insurance	do you have private health insurance? yes, no
*location	low SES, high SES
Habits	
*smoking	smoking,never smoked, have quit smoking
·alcohol	do you consume alcohol? yes,no
**exercise	do you exercise? yes,no
tvhours	hours of TV per day
**self-control	who usually checks blood pressure/blood sugar?: self-control, others
**visit frequency	when do you visit your health care provider?: to refill recipe, 3-6 months, when doctor calls, no visit
Health	
morbidity	existing conditions: DM, HT, both DM and HT
numberdrugs	number of medications taken regularly
under-control	do you think your condition is under control? : yes, no, I don't know
healthscore	self-assessed health score (1-10 scale)
*BMI	Body Mass Index (calculated using height and weight information)
diseaseduration	years since diagnosis of the disease
Health Service	
regularprovider	who is the regular health service provider: GP, private hospital, public hospital
diagnosiswho	who diagnosed the chronic condition? GP, specialist, ER
admission	any hospitalization in the last year:yes, no
ERvisit	any ER visit in the last year: yes, no
GPaccess	access time to GP: 1-2 days, up to a week, up to a month
hospital access	(access time to hospital): 1-2 days, up to a week, up to a month
diabetes training	yes, no
*vaccination	adult vaccination: yes, no
GP calls patient	Does GP make any calls to the patient? yes, no
asksmoking	Does GP ask the patient about smoking? yes, no
asksexercise	Does GP ask about exercise? yes, no
recommendsexercise	Does GP recommend exercise? yes, no
referral	Does GP refer to specialists: yes, no

The ones that are significant in Logistic Regression models for each category are indicated with Signif. codes: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

presents the distributions of age and PAM level in the study sample. Finally, blood pressure and blood sugar were measured, and HgA1c and blood pressure levels are noted from their health records with permission, when available. Unfortunately, the missing data amount in the health records was significant (57% of patients missed HgA1c levels, 39% of patients missed blood sugar levels in their health records); as a result, these features could not be investigated effectively. A preliminary data analysis was presented at the European Public Health Conference in 2019 [39].

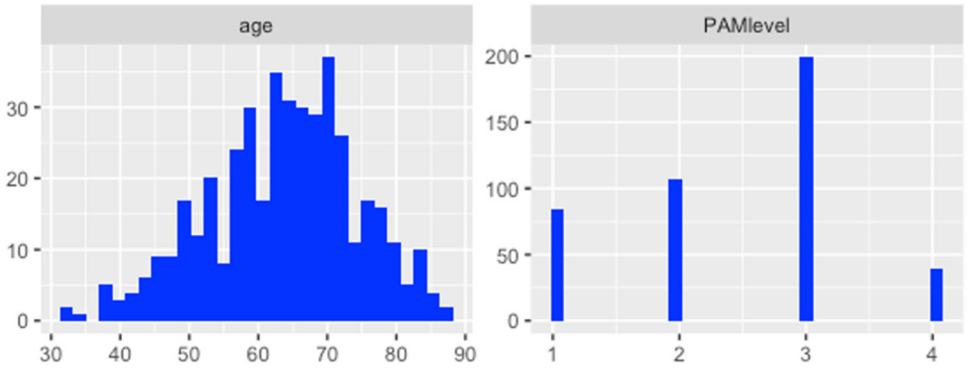

Fig. 1 Distribution of Age and PAM level in the study sample

3.3 Implementing classification algorithms for prediction of PAM

We aim to correctly identify Low PAM patients using machine learning approaches, making our problem a binary classification problem. This section first discusses data preprocessing techniques to make the survey data compatible with machine learning algorithms. Then, we introduce two feature selection approaches and elaborate on the employed machine learning algorithms. We later introduce our performance metrics and describe how we compute them.

Data Preprocessing Table 1 demonstrates that the collected survey data is mixed, i.e., we have both categorical and numerical data. Hence, the first issue in the data preprocessing part is to ensure that all categorical data is converted into a numerical representation. Towards this goal, we employ One-Hot-Encoding to transform the categorical variables into binary variables since most of the categorical variables do not present an ordinary relation among the values of the categorical features.[3]

Initial exploratory data analysis revealed that our dataset contains missing values. However, the machine learning algorithms we employ in this study assume that the data is represented via complete numerical matrices or arrays, which makes our dataset incompatible with these methods. In practice, two fundamental approaches tackle this problem: (1) Discard entire rows possessing missing values. (2) Impute missing values inferring the given part of the data. Frequent category imputation can be a useful approach for categorical data, whereas mean, median, or mode imputation can be utilized to tackle missing values in numerical data. Even though ignoring some rows comes at the price of losing conceivably valuable data, we employ this approach as our preliminary experiments showed that data imputation techniques lead to overfitting problems as we have a small number of participants. Nevertheless, we adopt a two-stage data-dropping policy to minimize data loss. First, we check the number of missing values for each feature. The feature is labeled inefficacious and deleted if the number exceeds a predetermined threshold. Otherwise, the rows with missing values for the relevant feature are deleted. We set the threshold value to 50, corresponding to more than 10% of the data. As a result, the two clinical outcomes, HgA1c and blood pressure levels from health records were deleted as the number of missing rows was very high.

Feature Selection The columns that survived after preprocessing are referred to as features. Feature selection refers to the techniques that preferably select the best subset of features to improve the capabilities of the machine learning algorithms by eliminating redundant and irrelevant features.

Ideally, one could test all possible combinations of the features, i.e., find the best set by a complete enumeration. However, it is not feasible to accomplish this for our problem, considering the number of features. Therefore, we employ two different feature selection approaches. The first is one of the traditional approaches called "Recursive Feature Elimination (RFE)." RFE is a wrapper-type feature selection algorithm. In other words, a machine learning algorithm is embedded into it to help select the features. It begins with all features and iteratively removes them by considering the given performance metric until the desired number of features remains. We employ this method with all the classification algorithms that we test.

Our second approach exploits domain-specific knowledge. We follow a two-step approach and use the four basic feature sets (demographics, habits, health, and health services) listed in Table 1. In the first step, we consider each group of features and select a subset of these features. At this step, we consider several approaches to eliminate features, which will be explained in more detail in Section 4.2. After extracting the selected features from each feature set, we search for the best combination of those sets in the second step. We employ the best subset selection method, i.e., evaluate all combinations of feature sets at this stage as we have only four feature sets. In addition, we also investigate other

[3] After encoding, 33 unique features correspond to 87 encoded features.

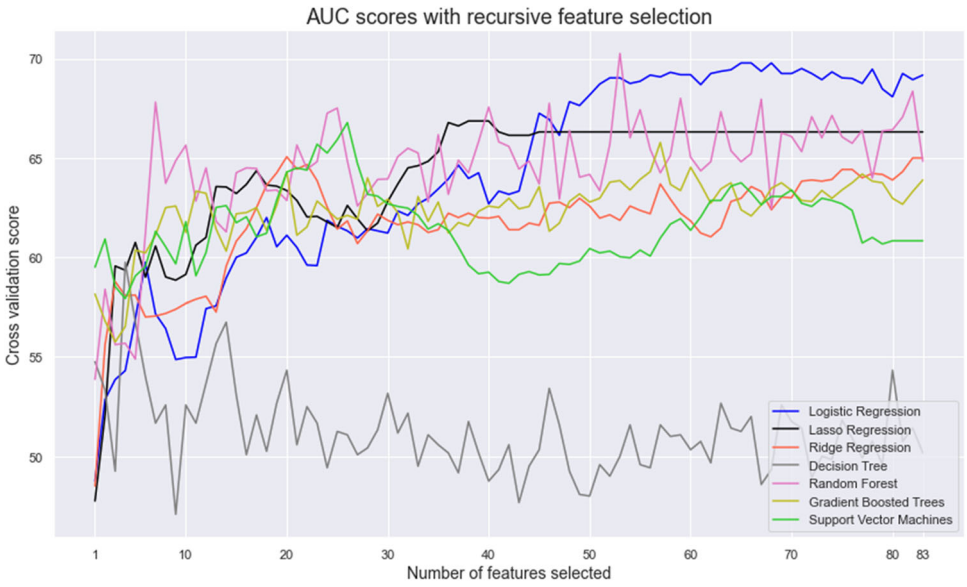

Fig. 2 The AUC scores achieved with Recursive Feature Selection for each algorithm and for different number of features selected

feature sets where we pick features based on our domain knowledge and discuss the performance effects of different features.

Machine Learning Algorithms We implement eight well-known machine learning approaches: (1) Logistic Regression, (2) Lasso Regression, (3) Ridge Regression, (4) Decision Trees (DT), (5) Random Forest (RF), (6) Gradient Boosted Trees (GBT), (7) Support Vector Machines (SVM), and (8) Neural Networks (NN). We implement each algorithm by using the Scikit-Learn library in Python programming language. All of the experiments are performed by the default values of Scikit-Learn for the parameters as an extensive grid search on hyperparameter values did not improve the performance significantly.

Performance Metric and Cross-Validation We employ the Area Under the Receiver Operator Characteristics (ROC) Curve (AUC) as the performance metric, which is a commonly used metric in the literature (cf. [37]).

A ROC curve plots True Positive Rate vs. False Positive Rate at different classification thresholds. The AUC measures the area below this curve; a perfect classifier would have an AUC of 1. Thus, a higher AUC is desirable. We apply a stratified k-fold cross-validation approach to reduce the variance in evaluating the performance of the algorithms, with k=50 folds, using the *StratifiedKFold* function from the Scikit-learn model selection library.

4 Results

In this section, we present results with two approaches for feature selection. First, in Section 4.1, we rely on data only and implement a recursive feature elimination algorithm. Second, in Section 4.2, we use our domain knowledge to pick features and present the performance of several alternative feature sets. This will allow us to comment on the effects of using domain knowledge on classification performance.

4.1 Classification with automated feature selection

In this section, we present our results from the Recursive Feature Elimination approach to feature selection, which is solely based on data and does not use domain knowledge. This performance will serve as a benchmark for the performance with using domain knowledge.

Figure 2 shows the evolution of the AUC scores as the number of features selected increases for each algorithm except neural networks.[4] Table 2 presents the maximum mean AUC score and the number of features selected to achieve that score for each algorithm, where the number of features before the one-hot-encoding is also reported, labeled as unique features.[5] Random Forest performs the best over-

[4] We have excluded neural networks as a recursive feature selection method in our study, specifically using the RFECV algorithm from the sklearn library. This is because RFECV requires the availability of 'coef' or 'feature importances' attributes, which are not applicable in the case of neural network implementations. Moreover, one of the fundamental concepts underlying neural networks is to enable the algorithm to learn feature importances in a black-box manner without the explicit need for feature selection.

[5] We count each categorical feature in the original survey as one unique feature while the classification algorithms use each level as one feature after one-hot-encoding. For example, education is a unique categorical feature, and a university degree is one of the features generated from this unique feature.

Table 2 Mean ROC AUC Performance with recursive feature selection algorithm, for low PAM=1 and 2

Method	Encoded Features	Unique Features	AUC
Random Forest	53	31	0.70
Logistic Regression	68	29	0.70
Lasso Regression	38	27	0.67
SVM	26	22	0.67
GBT	57	33	0.66
Ridge	20	13	0.65
DT	4	4	0.60

all, achieving AUC=0.70 with 31 unique features, which correspond to 53 features after one-hot-encoding. Logistic Regression performed equaivalently, using 29 unique features. As expected, LASSO regression uses fewer features. However, the prediction performance is significantly lower than the Logistic Regression.

4.2 Classifiction with feature selection using domain knowledge

Given a very large number of features relative to the data size, using domain knowledge may help reduce the feature set and improve performance. The features given in Table 1

Table 3 Feature sets used in machine learning algorithms

Feature Set	Features	AUC
(1) Main categories:		
Habits (H)	smoking, alcohol, exercise, self-control, visit frequency	0.67
Health (He)	morbidity,numberdrugs, under-control,	
	healthscore , BMI	0.68
Health Services (HS)	regularprovider, diagnosiswho, vaccination	
	GP calls patient	0.60
Demographics (Demo)	education, location	0.67
(2) Best Performing Combination of main categories:		
Health+Habit		0.72
Health+Demo		0.70
HS+Health+Habit		0.72
Habit+Health+Demo		0.72
Habit+HS+Demo		0.71
Demo+Habit+Health+HS		0.72
(3) Including self-control:		
Demo+self-control	education, location, self-control	0.68
Health+self-control	morbidity, numberdrugs,self-control	
	under-control, healthscore	0.69
HS+self-control	regularprovider, diagnosiswho, vaccination,	
Demo+Health+self-control	education, location, morbidity, numberdrugs	
	under-control, self-control	0.72
Habit without self-control	smoking, exercise, visit frequency	0.64
(4) Records Based		
Records 1	education, occupation, numberdrugs,morbidity	0.67
Records 2	education, occupation, numberdrugs,morbidity,BMI	0.66
Records 3	education, occupation, morbidity	0.68
Records 4	education, occupation, morbidity , BMI	0.68

AUC values are presented for the Logistic Regression Classifer

Table 4 Mean ROC AUC performance with alternative feature sets with different significance thresholds used in filtering features using Logistic Regression

	AUC							
	Features Included	LR	Lasso	Ridge	DT	RF	GBT	SVM	NN
habits $\alpha = 0.05$	smoking, exercise self-control, visit freq.	0.66	0.66	0.66	0.63	0.62	0.65	0.65	0.62
habits $\alpha = 0.01$	exercise self-control, visit freq.	0.64	0.65	0.65	0.63	0.65	0.66	0.64	0.65
health $\alpha = 0.05$	BMI	0.58	0.59	0.59	0.53	0.54	0.57	0.59	0.53
HS $\alpha = 0.1$	vaccination, GPcalls	0.57	0.58	0.58	0.57	0.57	0.57	0.58	0.57
HS $\alpha = 0.05$	vaccination	0.57	0.57	0.57	0.57	0.57	0.57	0.50	0.57

are grouped in four sets based on the type of information they provide. We first consider each set separately and try to find the best feature subset within each set to represent the given category of information. Towards this goal, we use several pre-processing approaches discussed in this section.

Feature Selection within Subsets We first fit Logistic Regression for each subset and include significant features at $\alpha = 0.1$ for subsets Demographics and Habits. More selective approaches with smaller threshold α values did not perform better for these categories, as seen in Table 4. For the subset Health Services and Health, we included features that are essential indications of health characteristics and patients' relations with the service providers, despite not being significant. We provide performance with different thresholds in Table 4, showing that this more inclusive approach for the subsets Health and Health Services improved performance.

We also consider correlations between features to avoid using highly correlated features. Table 5 and Fig. 3 provide the analysis of pairwise correlations for the numerical features. Interestingly, correlations are found to be minimal. Diabetes duration is weakly correlated with age and the number of drugs. We included only the number of drugs and not the diabetes duration in the feature subset related to Health.

For the categorical features, a chi-square test of independence is performed for each pair, and the pairs that are associated ($H0$ rejected at $p < 0.05$) are reported in Table 6. Despite being associated, education and location are included in the set Demographics. The performance with only one of these features was significantly lower (AUC=0.65 and 0.58, respectively, for education and location as single predictors). Therefore, these features are kept in the final feature set for the subset Demographics. Interestingly, gender was highly associated with many features, including education, occupation, and income, indicating a gender disparity in many socio-economic indicators. Including this feature in the subset did not improve the prediction performance for any combination of subsets (maximum AUC remained at 0.72); hence we left gender out of the feature set.

The features *under-control* and *morbidity* are associated with many others as well as with each other. However, removing either of these features reduced the prediction performance (AUC=0.63 vs. AUC=0.66, respectively, after removing these features). Therefore these features are also kept in the subset Health. Other correlated features, such as income, and occupation, are left out of the feature sets.

We thus finalize the feature subsets within each category (final sets are given in Table 3), and we test classification

Table 5 Pairwise Pearson correlation coefficients between numerical features

	age	diabetes duration	number drugs	screen hours	self healthscore	BMI	PAMlevel
age	1.00	0.30	0.18	0.11	0.15	-0.15	0.02
diabetesduration	0.30	1.00	0.29	0.03	0.00	-0.08	0.06
numberdrugs	0.18	0.29	1.00	0.00	-0.09	0.13	0.02
screenhours	0.11	0.03	0.00	1.00	-0.07	0.00	0.02
selfhealthscore	0.15	0.00	-0.09	-0.07	1.00	-0.10	0.09
BMI	-0.15	-0.08	0.13	0.00	-0.10	1.00	-0.17
PAMlevel	0.02	0.06	0.02	0.02	0.09	-0.17	1.00

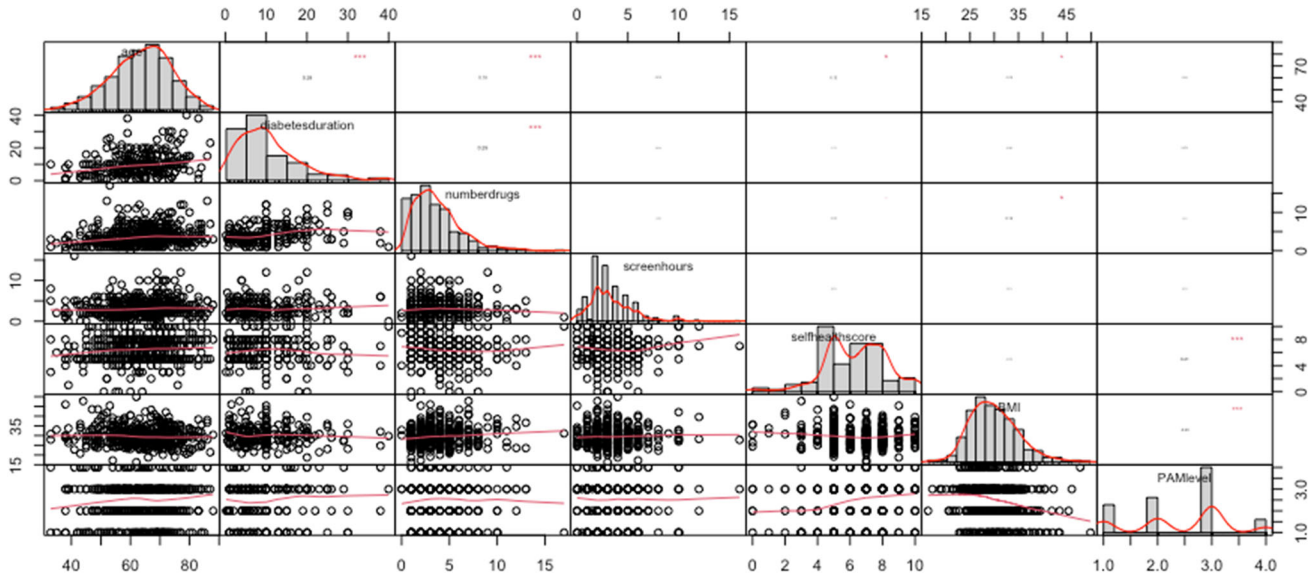

Fig. 3 Scatter Plots and Correlations for each numerical feature pair. Histograms for each variable is given on the diagonal

algorithms with different combinations of these feature subsets, as explained in the following.

4.2.1 Performance of algorithms for different feature sets

We consider different feature sets as follows. The first group takes each main subset only. Second, we test all possible combinations of these subsets. Third, we add one key feature (control-self) to each subset. Finally, we consider three new feature sets labeled as Records Based sets, considering the availability of information. Table 7 presents the mean AUC achieved with each algorithm and feature set tested. The 95% confidence intervals around these means are found using the sample of fifty accuracy levels found from the stratified k-fold procedure and are shown in Fig. 4 for the base feature sets but omitted in Table 7 for brevity.

Among the eight algorithms we implemented, Logistic Regression achieved the best performance for most feature sets. Logistic Regression is a traditional classification method, which also has the advantage of being interpretable. The coefficients of features in the Logistic Regression model result can be interpreted as feature importance. In a few cases, regression with Lasso and Ridge penalties slightly improved performance. The worst performing algorithm is Decision Trees, which produced the lowest AUC performance for all subsets. The remaining algorithms' performance was comparable to that of Logistic Regression. We think the relatively high number of features compared to the available data is one reason for the better performance of simpler linear models for prediction, as the risk of overfitting increases with numerous features in non-linear models [3, 14].

The highest AUC is 0.72 achieved with the Logistic Regression classifier. While these figures cannot be considered very high for a machine learning classification, they are comparable to the earlier studies in this area. For example, recent work [37] achieved a ROC of 0.684 to predict medication adherence with a Random Forest model, using a large dataset from Electronic Health Records, with features of demographic variables and previous adherence behavior.

Table 6 Features that are associated based on pairwise chi-square test of independence, at $\alpha = 0.05$

Feature	associated features ($p < 0.05$)
location	gpaccess, education, morbidity
gender	education, smoking, income, occupation, morbidity
	undercontrol, location
education	location, gender, education, income, occupation morbidity
income	diagnosiswho, gpaccess, education, occupation undercontrol
occupation	gender, diagnosiswho, education, income morbidity smoking, undercontrol
morbidity	location, diagnosiswho, education, occupation, undercontrol
smoking	gender, hospitalaccess, education, occupation, undercontrol
undercontrol	diagnosiswho, gpaccess, hospitalaccess, income, occupation, morbidity, smoking

A feature is listed when at least one of its levels is related with another feature

Table 7 Mean ROC AUC Performance with alternative feature sets formed using main subsets

	AUC							
	LR	Lasso	Ridge	DT	RF	GBT	SVM	NN
(1) Main Subsets:								
Habit	0.67	0.67	0.67	0.61	0.62	0.64	0.66	0.63
Health	0.68	0.68	0.67	0.53	0.58	0.60	0.66	0.62
HS	0.60	0.59	0.60	0.55	0.55	0.56	0.58	0.55
Demo	0.67	0.67	0.67	0.67	0.67	0.67	0.64	0.66
(2) Combinations of Subsets:								
Health+Habit	0.72	0.71	0.71	0.52	0.66	0.64	0.70	0.68
Habit+HS	0.67	0.65	0.67	0.55	0.61	0.68	0.65	0.66
Habit+Demo	0.69	0.68	0.69	0.60	0.63	0.65	0.69	0.67
Health+HS	0.68	0.68	0.67	0.52	0.58	0.59	0.66	0.66
Health+Demo	0.70	0.71	0.71	0.52	0.60	0.64	0.71	0.63
HS+Demo	0.68	0.68	0.68	0.59	0.61	0.65	0.69	0.63
HS+Health+Habit	0.72	0.69	0.70	0.55	0.66	0.65	0.68	0.64
Habit+Health+Demo	0.72	0.72	0.71	0.57	0.67	0.64	0.70	0.70
Habit+HS+Demo	0.68	0.67	0.68	0.59	0.61	0.70	0.70	0.69
Health+HS+Demo	0.71	0.70	0.69	0.55	0.64	0.65	0.71	0.70
Habit+Health+HS+Demo	0.72	0.70	0.70	0.59	0.68	0.65	0.70	0.68
(3) Including Self-Control:								
Demo+selfcontrol	0.68	0.68	0.68	0.67	0.67	0.67	0.66	0.67
Health+self control	0.69	0.70	0.69	0.54	0.62	0.65	0.67	0.63
Habits- self control	0.64	0.64	0.64	0.60	0.62	0.61	0.63	0.60
Health Service + self control	0.62	0.62	0.63	0.58	0.58	0.59	0.62	0.58
Demo+Health+self control	0.72	0.71	0.71	0.55	0.64	0.65	0.70	0.69
(4) Records Based:								
records1	0.67	0.67	0.67	0.63	0.63	0.67	0.66	0.64
records2	0.66	0.67	0.67	0.55	0.62	0.65	0.69	0.63
records 3	0.68	0.68	0.68	0.62	0.63	0.63	0.65	0.64
records 4	0.68	0.68	0.68	0.62	0.63	0.63	0.65	0.64
records 1 +self control	0.69	0.69	0.69	0.62	0.65	0.69	0.70	0.66
records2 + self control	0.71	0.70	0.71	0.55	0.67	0.68	0.69	0.67
records3 + self control	0.68	0.68	0.68	0.61	0.63	0.63	0.68	0.63
records4 +self control	0.71	0.69	0.70	0.55	0.64	0.64	0.70	0.65

4.2.2 Most useful features for prediction

The first groups of feature subsets consist of features related to a single dimension, such as Demographics, Health, or Habits. Among these, the highest AUC is achieved by using the *Health* features; therefore, we can assert that the most useful information to predict patient activation is the information about the current health status. However, the Demographics and Habits related information are almost equally useful. Using only a patient's education level and location (categorized as regions representing low SES or high SES), a Logistic Regression Model achieved an AUC of 0.67 with these two questions. In cases where Health-related information is unavailable, this basic demographics-related information can be quite useful.

The information set about the life-style and habits of the patient (*Habits*) also performs similarly, with a slightly lower AUC (0.67) compared to Health. Interestingly, the features related to health services, including operational factors, such as the information about the waiting times for healthcare service access and utilization of Emergency Rooms (ER) and inpatient services, do not help much for the prediction of activation in this data set when used alone. The most relevant health service-related features were the regular service provider, the provider doing the first diagnosis, vaccination status, and whether the GP calls the patient to check on them. The predictions are barely better than a random guess with an AUC level of only around 0.60 with these four features.

Fig. 4 Summary results for the base features sets

The second group of feature subsets combines different categories. The prediction performance improves by using Demographics and health-related features together (AUC=0.70 for the set Health+Demo). The highest AUC is achieved by using features related to Habits and Health (AUC=0.72 for the set Health+Habit). Adding features indicating Health Service use or Demographics to these does not hurt or improve prediction performance for the logistic regression classifier.

The third group of features investigates the importance of one particular feature, "self-control". The feature "self-control" is a binary variable indicating whether the patient checks their own blood sugar and blood pressure or lets someone else (such as a caretaker, relative, or pharmacist) do the control. This feature measures one aspect of the more general concept of "self-monitoring", which has been shown to be an essential component of self-management of chronic diseases [28]. Self-monitoring in the context of chronic illness has been defined as the patient undertaking one or more of the following activities (i) self-measurement of vital signs, symptoms, behavior, or psychological well-being; (ii) self-interpretation of this data; or (iii) self-adjustment of medication, treatment, lifestyle or help-seeking behavior as a result of self-awareness and/or self-interpretation [28].

Adding the self-control question to the two questions about education and location (a proxy for socio-economic status) increases the AUC of Logistic Regression to 0.68. Further, we add this question to the single-dimensional feature subsets, i.e., the feature sets in the first group. As the set Habits already include this feature, we remove it from Habits to see the relative effect with this subset. The third group in Table 3 presents the mean AUC values with these feature subsets, constructed with the addition of "self-control". A comparison of performance for feature sets that only differ in this feature (for example, set Demo vs. Demo+self-control) shows that this single feature increases the AUC by 0.01 to 0.02 for sets Demo, Health, and Health Services (HS), corresponding to an average 2% increase. For the HS subset, the difference in AUC is 0.04, corresponding to a 5% increase. Thus, we conclude that the "self-control" feature is an important feature, useful in predicting PAM level. When the information about Habits is unavailable, using only Demographics and Health-related features with the addition of self-control can achieve the best performance (the set Demo+Health+self-control with AUC=0.72).

Note that smoking, exercise habits, or self-control information may not always be available. In the fourth group of feature sets, we investigate the predictive power of feature sets that may be available in health records. We consider four such sets. *Records 1* only considers four features (education, occupation, number of medications, and morbidity) and achieves an AUC of 0.67. The *Records 2* adds BMI to the features but cannot improve performance. The *Records 3* leaves the number of drugs and BMI out while *Records 4* includes BMI. While these records-based feature sets cannot achieve good performance alone, adding self-control to the set Records 4 could increase the AUC to 0.71.

For Logistic Regression classifiers coefficient of the feature can be interpreted as the feature importance[6]. In Fig. 5,

[6] As all features (except for BMI and number drugs) are binary variables, we do not standardize the coefficients

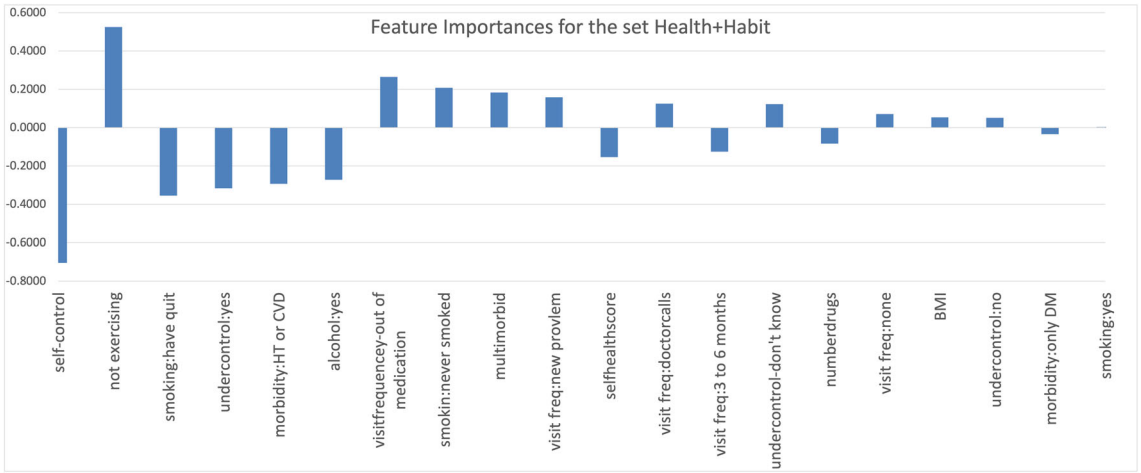

Fig. 5 Feature Importances for the LR classifier, averaged accross fifty folds of test-train split for the feature set Health+Habit

we plot the average value of the feature coefficients from the implementation of Logistic Regression classifier with StratifiedKfold procedure for the feature set "Health+Habit", which was the minimum sized feature set achieving maximum AUC. The features are sorted in decreasing absolute values. This figure also confirms that the *self-control* feature is the most important one, followed by smoking and exercise habits. We also provide the feature importances when the set Demographics is added to the features in Fig. 6. This set has a comparable prediction performance, and it is insightful to see the high importance of education and socio-economic status (represented by location) in this figure. These average coefficient values can be found in table form in the Appendix.

Comparing the performances of features selected using domain knowledge and features selected by the recursive algorithm, we can conclude that using domain knowledge performs better, with a two percentage point difference in the mean AUC. In addition, the domain knowledge-based hierarchical approach presented in Section 4.2 achieved this performance with only ten unique features, less than half of

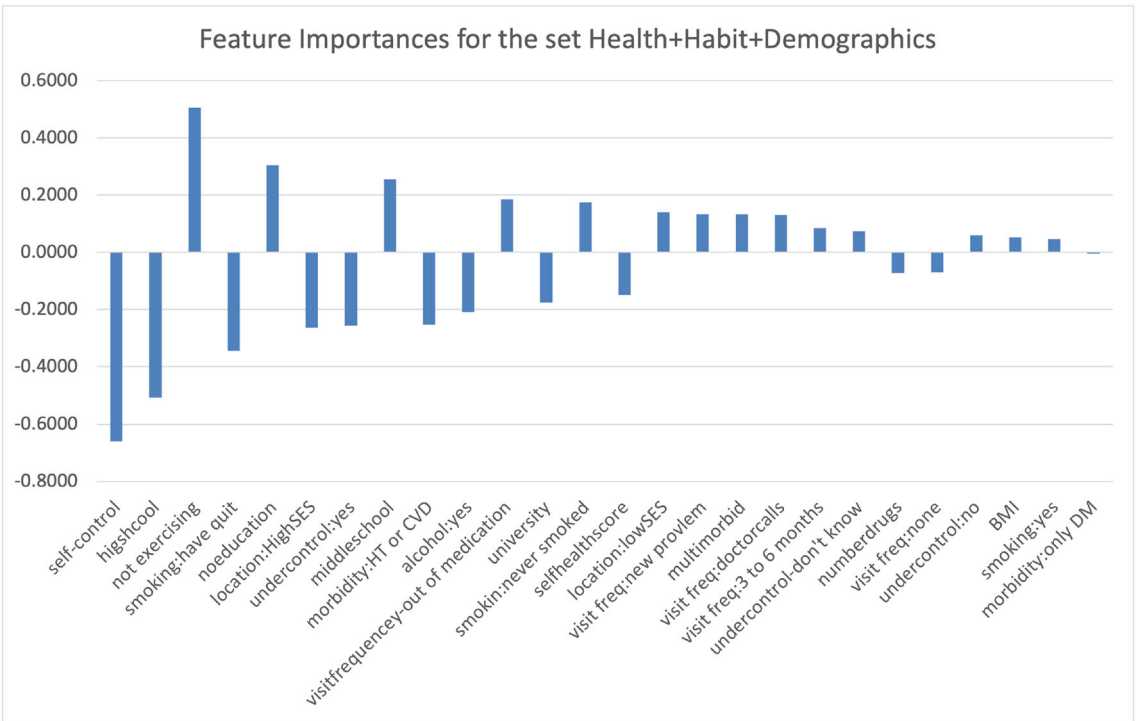

Fig. 6 Feature Importances for the LR classifier, averaged accross fifty folds of test-train split for the feature set Health+Habit+Demographics

the Recursive Feature Elimination algorithm. This result provides evidence about the benefits of using domain knowledge in machine learning.

4.2.3 Further refinements for features

A further improvement in the performance of the selected features could be achieved by adding interactions of features to a good-performing feature set. The Interactions variables can model moderation effects, which are very common in many practical settings. We identified a set of possible interactions in this domain. For example, the meaning of self-control can be different for an educated and an uneducated patient. If a highly educated patient does not do self-control, it may be a stronger signal of low activation than the case of an uneducated patient not doing self-control. Thus we included the interaction "education x self-control". In addition, we added the following interactions: notexercisingxBMI, multimorbidxsmoking, multimorbidxalcohol, and livealonexself-control. We repeated the test of the algorithms using the Stratified 50-fold procedure with this more extensive feature set: the prediction performance of regression methods slightly improved with these interaction terms. The highest performance was obtained when only the interactions multimorbidxalcohol, noeducationxselfcontrol, universityxselfcontrol, and notexercisingxBMI were added, where AUC=0.726 by Logistic Regression. We think this slight improvement does not justify the increased complexity of the model; therefore, in the next section, we will present a regression model without interaction terms.

4.3 A logistic regression model for identification of low PAM patients

In this section, we fit a logistic regression model for the whole data set using one of the feature sets with the highest AUC, Demo+Health+Habits. We include the set Demographics as this adds only two features, which are also very easy to obtain and deemed important in the literature. The coefficients for other logistic regression classifiers with feature subsets that perform similarly are given in the Appendix. We have two purposes here; to illustrate developing a practical tool and to discuss associations of feature values with the probability of having low PAM.

Table 15 presents the coefficients for the logistic regression model. This model takes the LowPAM variable as the dependent variable, which is set to 1 if the PAM level equals 1 or 2. A prediction algorithm can be based on such a model. The algorithm predicts the probability of having low PAM using this model and then predicts the patient to have low PAM if this probability is greater than 0.50. Otherwise, the patient is classified as having high PAM. The sign of the coefficients shows the direction of the effect of each variable used in regression[7].

Here coefficient magnitudes can be interpreted as the importance of the features. We observe that not having formal education (noeducation), or having little formal education (middle school) increases the risk of low PAM. At the same time, those who do their checks by themselves are at lower risk. Accordingly, education, self-control, and exercise emerge as the most important features, which are also statistically significant.

Another interesting result is that patients who quit smoking are less likely to have low PAM than people who never smoked. This is also intuitive. Quitting smoking requires a lot of effort and willpower; thus, a person who did quit smoking is very likely to be a person who adopts healthy behavior, implying a highly activated patient. Other significant features are exercise habits and BMI; not exercising and higher BMI are predictors of low PAM. Interestingly, consuming alcohol predicts a lower risk of low PAM, although not significant. This is probably due to its association with higher Socio-Economic status in the context of Turkey, an interesting result that may not be valid in other countries. Other predictors of low PAM are using fewer medications and having multimorbidities (both Diabetes Mellitus and Hypertension or cardiovascular disease in our case). Finally, being able to answer the question "is your condition under control?" as "yes" or "no" predicts a lower likelihood of low PAM, as indicated by the negative coefficients in Table 8. The reference level is "I don't know", which predicts higher likelihood of low PAM, compared to a yes or no answer.

The formula below can be used to calculate the probability of low PAM for a given patient. Let β_i be the coefficient of the feature i, where the features are numbered in Table 8, and let x_i be the value of the feature for the patient. Then,

$$z = \Sigma_i \beta_i x_i \tag{1}$$

$$P(low\,PAM) = \frac{1}{1+e^{-z}}. \tag{2}$$

Figure 7 shows the prediction performance of the above algorithm for a random $20\% - 80\%$ test-train split of the data, with 76 patients in the test set. The left panel shows the Receiver Operating Characteristics Curve, plotting the true positive rate of the algorithm as a function of the false positive rate. The different points on the plot represent the performance of the algorithm in the test set for different thresholds

[7] Note that the model presented in Table 8 uses dummy coding for the variables, dropping one category level for each categorical variable. The one-hot encoding procedure used in the cross-validation classification algorithm in the previous sections has not dropped levels for categories with more than two levels. Thus the number of features in Table 8 is less than the number of features in Fig. 6. Nevertheless, the main results are consistent with each other.

Table 8 Coefficients of the LR Model for the feature set D+H+He

Feature no	Feature	Estimate	Std.Error	z	$Pr(>\|z\|)$
0	(Intercept)	-1.747	1.059	-1.649	0.099
1	education-middleschool	0.935	0.330	2.833	0.005**
2	education-noeducation	1.312	0.501	2.621	0.009**
3	education-university	0.549	0.403	1.361	0.173
4	location-low SES	0.359	0.247	1.452	0.146
5	selfhealthscore	-0.112	0.065	-1.719	0.086 .
6	numberdrugs	-0.072	0.055	-1.308	0.191
7	morbidity-multimorbid	0.445	0.323	1.379	0.168
8	morbidity-DM	0.251	0.511	0.491	0.623
9	undercontrol-yes	-0.335	0.538	-0.623	0.533
10	undercontrol-no	-0.066	0.625	-0.105	0.916
11	BMI	0.069	0.025	2.794	0.005
12	smoking-never	-0.083	0.338	-0.245	0.807
13	smoking-have quit	-0.652	0.357	-1.826	0.068 .
14	alcohol-yes	-0.353	0.458	-0.770	0.441
15	exercise-no	0.676	0.259	2.607	0.009**
16	self-control-yes	-0.853	0.281	-3.041	0.002**
17	drvisit-outofmedication	0.291	0.268	1.086	0.277

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

used for classification. Note that we use the default threshold, 0.50, which corresponds to the blue dot on the plot. For this particular data split, the AUC is found as 0.744, and the accuracy is 0.67. [8]

The right panel of Fig. 7 shows the confusion matrix for this prediction experiment. Among the 76 test patients, 34 were actually low PAM patients, indicated with a true label equal to 1 (lowPAM=1). Twenty-one of these patients were classified correctly, indicating a sensitivity of 21/34=0.62. Out of 42 high PAM patients, 30 were predicted correctly, indicating specificity of 30/42=0.72. Overall prediction accuracy is then (30+21)/76=0.67. We note that the algorithm has a lower false positive rate (1-0.72=0.28) than the false negative rate (1-0.62=0.38).

4.4 Discussion

We used primary data collected via in-person surveys with the patients in the target group. The survey provided us with a wide range of features related to the demographics, health conditions, and lifestyle of patients and their utilization of health services. This unique data set allows us to comment on the relative importance of different feature categories.

We can summarize our findings as follows: (1) Logistic Regression method was the most successful among algorithms tested in our experiments; (2) The most useful characteristics that help to predict PAM level are related to health status (AUC=0.68). The information about education and Socio-Economic Status and information about habits alone can be almost equally useful (AUC=0.67). (3) Features that are likely available in electronic health records, such as education, occupation, morbidity, and BMI, can achieve similar prediction performance (AUC=0.68). (4) To improve the prediction performance further, having information about the patient's habits and attitudes is essential. Combining health and habit-related information achieves AUC=0.72. In particular, information on smoking and exercise habits proves useful. These observations suggest that patient activation is a multi-dimensional concept, and different types of information are needed to predict high-risk patients.

The question "Who usually checks your blood pressure/blood sugar?" emerged as a helpful question to ask a patient with Diabetes Mellitus and/or Hypertension. This question measures a minimum level of self-monitoring. A patient who does not do self-control has a higher risk of having low PAM. One drawback of such a question for prediction is that regular checks for blood sugar and blood pressure require tools, and less privileged patients may not have the means to do their own regular checks. Hence, it would also be important to understand the reasoning behind the answer to this question in practice. Another useful question is "Do you think your condition is under control?". Patients with the response "I don't know" are at higher risk of low PAM.

Interestingly, the features related to health services, including operational factors, such as the healthcare ser-

[8] Note that the formula used in this prediction is not exactly the same as that given above; the model in Table 15 is fitted to the whole data, while the below figures use a model fitted to the training set. However, to use this model for prediction in a new data set, the whole study data should be used, which results in the model given in Table 15.

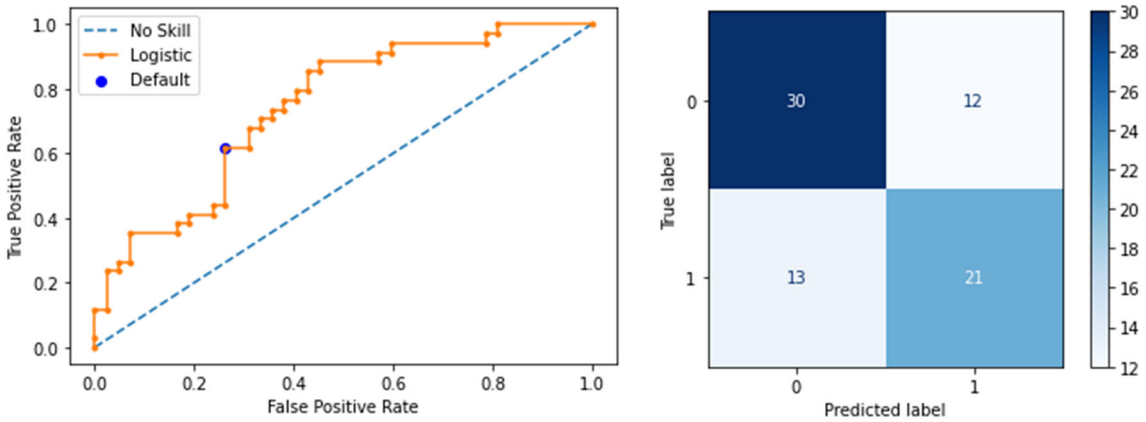

Fig. 7 Receiver Operating Characteristics Curve (left) and Confusion Matrix (right) for a random test-train set split of the data

vice access and utilization of Emergency Rooms (ER) and inpatient services, do not help much for the prediction of activation in this data set. One would expect that high PAM patients use the ER and inpatient services less frequently since their condition would more likely be in control, and they would see a doctor before they need an ER visit. However, our data do not lend support to this hypothesis. The survey question did not specify the reason for the ER visit, which may be a reason for the absence of this relation. In addition, Turkey is one of the countries with the highest number of ER visits per capita [1][9]. Moreover, although in the previous literature, a positive effect of primary healthcare service characteristics such as ease of access and perception of time spent with the physician on patient engagement was reported [32, 51], the access time was not an important feature for classification in our experiments. This may be due to the low variability and relatively short access times to see a GP. In future work, a larger dataset with more variability in primary healthcare characteristics can be investigated.

Despite the rich feature set we worked with, the average prediction performance was not very high. The low predictive power shows the difficulty of predicting patient activation, which is a concept related to patients' motivation, skills, and confidence in managing their health. Using only health records and objective information about patients may not be enough to predict PAM levels; having some more personal information about their lifestyle and habits is essential for this purpose. Nevertheless, our analysis highlights some interesting insights regarding high-risk patients. It also provides two questions that may be useful to identify high-risk patients, which clinicians and policymakers can use in practice.

[9] Based on data for the January-October period of 2017, the total number of emergency service examinations is 84,545,429, which is more than the total population of Turkey.

5 Extension: classification with different definitions of low PAM

In this study, we defined PAM levels 1 and 2 as low PAM as these patients are designated as needing more support in the UK. However, a patient needs more support as the PAM level decreases, and it is possible to apply our approach with other definitions for low PAM. In this section, we provide analysis for different cut-off PAM levels to define low PAM levels.

5.1 Analysis for low PAM=PAM level 1

In the data set, 20% of patients had PAM Level equal to 1, representing the patients that needed the most help. We aim to identify these patients in this section by defining low PAM as a PAM level equal to 1. Tables 9 and 10 provide the prediction performance of algorithms with automated feature selection, and with the feature sets selected previously in Section 4.2. In this case, Random Forest regression achieved the best performance with recursive feature elimination (AUC=0.70), using only eleven features. These features are as follows: occupation, smoking, exercise, ERvisit, self-control, visit frequency, age, numberdrugs, tvhours, selfhealthscore, and BMI. When domain knowledge is used, the Logistic Regression model achieves AUC=0.69 with only five features (subset Habit). Although its AUC performance is slighlty worse, Logistic Regression has the advantage of interpretability, therefore we can conclude using domain knowledge brings an advantage in this case as well.

5.2 Analysis for low PAM=PAM level 1,2 or 3

In the data set only 9% of patients had PAM level equal to 4, i.e., the data set is highly imbalanced. In this section, we consider these as high PAM, and those with PAM levels less than 4 are defined as low PAM. Note that this low PAM definition may not be viable in practice, and the classifica-

Table 9 Mean ROC AUC Performance with recursive feature selection algorithm, for low PAM=PAM Level 1

Method	Encoded Features	AUC
Random Forest	11	0.70
Logistic Regression	77	0.67
Lasso Regression	23	0.64
SVM	15	0.62
Ridge Regression	49	0.61
Gradient Boosted Trees	4	0.60
Decision Tree	7	0.56

Table 11 Mean ROC AUC Performance with recursive feature selection algorithm, for low PAM=PAM Level 1, 2, or 3

Method	Number of Features	AUC
lasso	11	0.81
ridge	30	0.79
lr	18	0.76
lasso	27	0.72
gbt	17	0.70
rf	55	0.70
dt	18	0.61

tion algorithms' performance may not be reliable due to the imbalance [9, 26]. Nevertheless, for the sake of completeness, we present the analysis for this data set in this section.

The performance of algorithms with automated feature selection, and with the feature subsets from Section 4.2 are presented in Tables 11 and 12, respectively. Interestingly, with two features defining the subset Demographics, education and location, the mean AUC is 0.73, and adding Habits related information increases the AUC to 0.74, the highest we could find in this study with the domain based features. The recursive feature selection algorithm performed significantly better in this case, achieving AUC=0.81 with lasso regression.

6 Conclusions

In conclusion, in this paper, we show that machine learning can be implemented to classify patients as having low activation or not. In addition, we made observations on the attributes that would be most useful for predicting the probability of patients having low PAM levels, which is correlated with their adherence behavior and health outcomes. Having an algorithm for prediction with a small subset of features may make it easier for the physician to allocate time and consider the PAM prediction in practice. This may be particularly valuable for overloaded practices and public health purposes.

We collected a rich data set from patients with chronic diseases, Hypertension, Cardiovascular Disease or Diabetes Mellitus Type II. The broad range of features collected allowed us to investigate the use of different information sets for prediction purposes. We used two approaches to future selection: a data-based approach with no domain knowledge and a hierarchical approach that groups futures based on contexts and picks useful features. Comparing these two approaches showed that using domain knowledge can achieve superior performance with fewer features.

Table 10 Mean ROC AUC Performance with subset combinations for each algorithm, for low PAM=Pam Level 1

	LR	Lasso	Ridge	DT	RF	GBT	SVM	NN
Habit	0.69	0.69	0.69	0.64	0.64	0.65	0.55	0.64
Health	0.59	0.62	0.64	0.51	0.53	0.55	0.58	0.50
HS	0.57	0.56	0.57	0.52	0.51	0.49	0.60	0.53
Demo	0.58	0.58	0.57	0.53	0.53	0.53	0.39	0.49
Health+Habit	0.69	0.68	0.68	0.51	0.61	0.64	0.69	0.55
Habit+HS	0.66	0.67	0.67	0.58	0.60	0.66	0.59	0.58
Habit+Demo	0.68	0.68	0.70	0.54	0.60	0.64	0.56	0.64
Health+HS	0.62	0.63	0.62	0.54	0.55	0.56	0.55	0.60
Health+Demo	0.62	0.66	0.67	0.50	0.54	0.56	0.56	0.58
HS+Demo	0.58	0.58	0.58	0.56	0.57	0.55	0.45	0.56
HS+Health+Habit	0.68	0.66	0.66	0.52	0.61	0.60	0.63	0.66
Habit+Health+Demo	0.68	0.68	0.67	0.51	0.60	0.60	0.61	0.66
Habit+HS+Demo	0.65	0.66	0.65	0.55	0.63	0.64	0.54	0.64
Health+HS+Demo	0.63	0.65	0.63	0.55	0.63	0.54	0.60	0.60
Habit+Health+HS+Demo	0.68	0.66	0.65	0.51	0.59	0.54	0.66	0.61

Table 12 Mean ROC AUC Performance with subset combinations for each algorithm, for low PAM=PAM Level 1, 2, or 3

	LR	Lasso	Ridge	DT	RF	GBT	SVM	NN
Habit	0.69	0.67	0.69	0.67	0.68	0.69	0.62	0.68
Health	0.52	0.54	0.49	0.48	0.46	0.53	0.50	0.48
HS	0.61	0.57	0.61	0.48	0.54	0.52	0.49	0.48
Demo	0.73	0.72	0.73	0.65	0.65	0.68	0.49	0.67
Health+Habit	0.61	0.64	0.61	0.49	0.48	0.55	0.61	0.61
Habit+HS	0.69	0.69	0.67	0.53	0.61	0.60	0.61	0.63
Habit+Demo	0.74	0.75	0.75	0.59	0.63	0.69	0.69	0.65
Health+HS	0.55	0.57	0.52	0.49	0.53	0.46	0.52	0.54
Health+Demo	0.68	0.70	0.69	0.54	0.65	0.65	0.64	0.65
HS+Demo	0.71	0.72	0.72	0.54	0.63	0.70	0.56	0.66
HS+Health+Habit	0.59	0.65	0.60	0.51	0.55	0.54	0.59	0.64
Habit+Health+Demo	0.71	0.72	0.72	0.48	0.65	0.68	0.72	0.70
Habit+HS+Demo	0.73	0.76	0.74	0.55	0.57	0.68	0.69	0.71
Health+HS+Demo	0.67	0.69	0.68	0.62	0.65	0.69	0.66	0.68
Habit+Health+HS+Demo	0.70	0.72	0.70	0.54	0.65	0.62	0.56	0.68

We also comment on the most commonly used features and how much prediction performance can be improved by adding other attributes. In addition to the widely used demographic features related to health behavior, we identify an important feature that significantly improves predictions. The question of who is doing the regular daily checks of a patient is a strong predictor of the patient's activation level. A more active, engaged patient would do their daily health controls.

This finding may also give insights to improve patient engagement, although we cannot claim causality for these results. It may imply that encouraging a patient to do their own checks and focusing on teaching them how to do the checks may help them be more aware of their disease, manage it, and feel more confident about it. In future research, such an intervention can be tested.

This paper focused on a classification problem, defining high-risk patients as those with a PAM level lower than a threshold. We considered PAM levels 1 and 2 as low PAM patients and presented an extension of the analysis with different definitions. A classification approach may be advantageous as it provides a simple and direct definition of who is high risk. However, if prediction at the granularity of the PAM level is desired, either ordinal regression could be applied, or iterative use of classification algorithms could be implemented in future work.

This research also has some limitations, and all the results should be interpreted in light of these limitations. Data on clinical outcomes, which are objective indicators of chronic disease management and patient engagement, could not be used due to their incompleteness (such as blood sugar HbA1c and Blood Pressure level). The data size was limited relative to the number of features; this may have prohibited achieving a higher prediction performance. Furthermore, the data were collected from patients attending the primary care clinics, which may have created a selection bias. The lack of findings about the effect of health service access and healthcare utilization may be due to this selection bias. New data can be collected to test these models with different samples in future work.

Appendix

PAM Scale Questions

The following questions are included in the PAM, developed by [18]. Due to the proprietary nature of the instrument, scoring scales are not included:

1. When all is said and done, I am the person who is responsible for managing my health condition.
2. Taking an active role in my own health care is the most important factor in determining my health and ability to function.
3. I am confident that I can take actions that will help prevent or minimize some symptoms or problems associated with my health condition.
4. I know what each of my prescribed medications do.
5. I am confident that I can tell when I need to go get medical care and when I can handle a health problem myself.
6. I am confident I can tell my health care provider concerns I have even when he or she does not ask.
7. I am confident that I can follow through on medical treatments I need to do at home.
8. I understand the nature and causes of my health condition(s).

9. I know the different medical treatment options available for my health condition.
10. I have been able to maintain the lifestyle changes for my health that I have made.
11. I know how to prevent further problems with my health condition.
12. I am confident I can figure out solutions when new situations or problems arise with my health condition.
13. I am confident that I can maintain lifestyle changes like diet and exercise even during times of stress.

Additional Tables

Table 13 Summary statistics for the features in the data set

DEMOGRAPHICS			HEALTH		
	mean	std		mean	std
female	0.64		morbifity-onlyHT	0.253	
education-highschool	0.203		morbidity-multimorbid	0.674	
education-middleschool	0.501		morbidity-onlyDM	0.073	
noeducation	0.122		undercontrol-yes	0.823	
education-university	0.174		undercontrol-no	0.122	
occupation-notworkingoutside	0.37		undercontrol-don'tknow	0.055	
occupation-retired	0.538		selfhealthscore	6.422	1.927
occupation-working	0.092		numberdrugs	3.851	2.512
location-LowSEs	0.439		BMI	29.839	5.062
age	63.581	10.985			
privateinsurance	0.069				
livealone	0.132				
HABITS			HEALTH SERVICE		
	mean	std		mean	
smoking-yes	0.182		healthprovider-FHC	0.531	
smoking-neversmoked	0.489		HCprovider-publichospital	0.506	
smoking-havequit	0.329		HCprovider-universityhospital	0.032	
alcohol	0.09		HCprovider-privatehospital	0.115	
exercise-no	0.314		Hcprovider-privateclinic	0.010	
self-control	0.733		diagnosiswho-GP	0.078	
drvisit-outofmedication	0.695		diagnosiswhor-other	0.010	
drvisit-newproblem	0.374		diagnosiswho-specialist	0.912	
drvisit-threesixmonth	0.333		gpaccess-oneday	0.973	
drvisit-doctorcalls	0.05		gpaccess-1week-1month	0.002	
drvisit-nochecks	0.052		gpaccess-1-2days	0.020	
screenhours	3.37	2.182	gpaccess-3-7 days	0.000	
			hospitalaccess-morethan1month	0.042	
			hospitalaccess-lessthan1 day	0.196	
			hospitalaccess-1weekto1month	0.516	
			admission	0.186	
			ER visit	0.396	
			vaccination	0.303	

Table 14 Feature importances for the LR Classifier for the feature sets Habits+Health+Demo and Habits+Health, Averaged over fifty cross validation folds

Feature	Importance Value	Feature	Importance Value
self-control	-0.6593	self-control	-0.7043
higshcool	-0.5064	not exercising	0.5258
not exercising	0.5061	smoking:have quit	-0.3549
smoking:have quit	-0.3442	undercontrol:yes	-0.3172
noeducation	0.3046	morbidity:HT or CVD	-0.2932
location:HighSES	-0.2642	alcohol:yes	-0.2727
undercontrol:yes	-0.2569	visitfrequencey-out of medication	0.2647
middleschool	0.2544	smokin:never smoked	0.2078
morbidity:HT or CVD	-0.2520	multimorbid	0.1834
alcohol:yes	-0.2084	visit freq:new provlem	0.1586
visitfrequencey-out of medication	0.1849	selfhealthscore	-0.1541
university	-0.1766	visit freq:doctorcalls	0.1252
smokin:never smoked	0.1743	visit freq:3 to 6 months	-0.1247
selfhealthscore	-0.1487	undercontrol-don't know	0.1223
location:lowSES	0.1401	numberdrugs	-0.0829
visit freq:new provlem	0.1325	visit freq:none	0.0713
multimorbid	0.1324	BMI	0.0532
visit freq:doctorcalls	0.1302	undercontrol:no	0.0512
visit freq:3 to 6 months	0.0842	morbidity:only DM	-0.0338
undercontrol-don't know	0.0731	smoking:yes	0.0034
numberdrugs	-0.0729		
visit freq:none	-0.0696		
undercontrol:no	0.0596		
BMI	0.0517		
smoking:yes	0.0458		
morbidity:only DM	-0.0045		

Table 15 Coefficients of the LR Model for the feature set Habits+Health

Feature	Coefficient	Std. Error	z value	$Pr(>\|z\|)$
(Intercept)	-1.097	1.026	-1.069	0.285
selfhealthscore	-0.107	0.065	-1.654	0.098 .
numberdrugs	-0.088	0.055	-1.602	0.109
morbidity:multimorbid	0.582	0.316	1.839	0.066 .
morbidity:only DM	0.306	0.506	0.604	0.546
undercontrolevet	-0.709	0.515	-1.376	0.169
undercontrol:no	-0.445	0.606	-0.735	0.463
BMI	0.073	0.024	2.989	0.003 **
smoking:never smoked	0.017	0.330	0.051	0.959
smoking:have quit	-0.583	0.352	-1.655	0.098 .
alcohol:yes	-0.554	0.449	-1.233	0.218
notexercising	0.745	0.254	2.931	0.003 **
self-control	-0.930	0.275	-3.387	0.001 ***
visitfreq:outofmedication	0.588	0.282	2.084	0.037 *
visitfreq:newproblem	0.286	0.249	1.150	0.250
visitfreq:doctorcalls	0.642	0.535	1.200	0.230
visitfreq:nochecks	0.783	0.598	1.310	0.190

Signif. codes: 0.000 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 16 Logistic Regression Model for Set Demo+Health+self-control

	Estimate	Std.Error	z value	$Pr(>\|z\|)$
(Intercept)	-1.512	1.029	-1.470	0.142
education-middleschool	1.037	0.319	3.251	0.001**
education-noeducation	1.475	0.481	3.070	0.002**
education-university	0.500	0.390	1.281	0.200
location-low SES	0.386	0.239	1.615	0.106
selfhealthscore	-0.124	0.064	-1.927	0.054 .
numberdrugs	-0.081	0.054	-1.487	0.137
morbiditymultimorbid	0.457	0.313	1.460	0.144
morbidity-only DM	0.350	0.497	0.704	0.481
undercontrol-yes	-0.342	0.529	-0.646	0.518
undercontrol-no	0.108	0.613	0.177	0.860
BMI	0.067	0.024	2.789	0.005**
self-control	-0.919	0.273	-3.370	0.001***

AIC: 469.88
Signif. codes: 0.000 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 17 Logistic Regression Model for Set Demo+Health+Habit+Health Service

	Estimate	Std.Error	z value	$Pr(>\|z\|)$
(Intercept)	-1.153	1.136	-1.015	0.310
education-middleschool	0.867	0.337	2.576	0.010**
education-noeducation	1.322	0.512	2.584	0.010**
education-university	0.582	0.411	1.414	0.157
location-lowSES	0.403	0.253	1.595	0.111
selfhealthscore	-0.114	0.066	-1.721	0.085
numberdrugs	-0.051	0.057	-0.898	0.369
morbidity-multimorbid	0.510	0.333	1.529	0.126
morbidity-onlyDM	0.411	0.521	0.788	0.431
undercontrol-yes	-0.274	0.552	-0.496	0.620
undercontrol-no	-0.022	0.640	-0.034	0.973
BMI	0.071	0.025	2.787	0.005**
smoking-never smoked	-0.149	0.341	-0.437	0.662
smoking-have quit	-0.623	0.360	-1.731	0.083.
alcohol-yes	-0.421	0.463	-0.909	0.364
notexercising	0.676	0.262	2.576	0.010**
self-control	-0.838	0.285	-2.936	0.003**
drvisit-outofmedication	0.272	0.271	1.003	0.316
diagnosiswho-ER	0.626	1.671	0.374	0.708
diagnosiswho-specialist	-0.649	0.461	-1.408	0.159
privatehospital-yes	-0.072	0.361	-0.199	0.842
vaccination-yes	-0.449	0.273	-1.647	0.100.
gpcalls-yes	-0.380	0.473	-0.803	0.422

AIC: 469.88
Signif. codes: 0.000 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Acknowledgements We sincerely thank the reviewers for their constructive comments that significantly improved this paper. We are grateful to the AXA Research Fund for the financial support provided through the AXA Award granted to the second author.

Declarations

Conflict of Interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethical standard The study was approved by the Ethis Committee of Marmara University. The data is available on request.

References

1. Turkish Ministry of Health Summary Statistics for Emergency Room Use in (2017). https://khgmistatistikdb.saglik.gov.tr/Eklenti/23496/0/2017-ocak-ekim-donemi-acil-servis-verileri2pdf.pdf. Accessed: 2021-07-04
2. Allen LN, Feigl AB (2017) What's in a name? a call to reframe noncommunicable diseases. Lancet Global Health 5(2):e129–e130
3. Alpaydin E (2020) Introduction to Machine Learning (Adaptive Computation and Machine Learning Series). The MIT Press, 4 edition
4. Awan SE, Sohel F, Sanfilippo FM, Bennamoun M, Dwivedi G (2018) Machine learning in heart failure: ready for prime time. Curr Opinion Cardiol 33(2):190–195
5. Bonnell LN, Littenberg B, Wshah SR, Rose GL (2020) A machine learning approach to identification of unhealthy drinking. J Am Board Family Med 33(3):397–406
6. Carman KL, Dardess P, Maurer M, Sofaer S, Adams K, Bechtel C, Sweeney J (2013) Patient and family engagement: a framework for understanding the elements and developing interventions and policies. Health Affairs 32(2):223–231
7. Chew S, Brewster L, Tarrant C, Martin G, Armstrong N (2018) Fidelity or flexibility: an ethnographic study of the implementation and use of the patient activation measure. Patient Educ Counsel 101(5):932–937
8. Dentzer S (2013) Rx for the 'blockbuster drug' of patient engagement
9. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) Learning from imbalanced data sets, vol 10. Springer
10. Gao J, Arden M, Hoo ZH, Wildman M (2019) Understanding patient activation and adherence to nebuliser treatment in adults with cystic fibrosis: responses to the uk version of pam-13 and a think aloud study. BMC Health Serv Res 19(1):1–12
11. Graffigna G, Barello S, Bonanomi A, Lozza E (2015) Measuring patient engagement: development and psychometric properties of the patient health engagement (phe) scale. Front Psychol 6:274
12. Greene J, Hibbard JH (2012) Why does patient activation matter? an examination of the relationships between patient activation and health-related outcomes. J General Int Med 27(5):520–526

13. Greene J, Hibbard JH, Sacks R, Overton V, Parrotta CD (2015) When patient activation levels change, health outcomes and costs change, too. Health Affairs 34(3):431–437 (**PMID: 25732493**)
14. Grinsztajn L, Oyallon E, Varoquaux G (2022) Why do tree-based models still outperform deep learning on tabular data? arXiv:2207.08815
15. Hibbard JH, Greene J (2013) What the evidence shows about patient activation: better health outcomes and care experiences; fewer data on costs. Health affairs 32(2):207–214
16. Hibbard JH, Greene J, Overton V (2013) Patients with lower activation associated with higher costs; delivery systems should know their patients' 'scores'. Health Affairs 32(2):216–222
17. Hibbard JH, Greene J, Sacks R, Overton V, Parrotta CD (2016) Adding a measure of patient self-management capability to risk assessment can improve prediction of high costs. Health Affairs 35(3):489–494
18. Hibbard JH, Mahoney ER, Stockard J, Tusler M (2005) Development and testing of a short form of the patient activation measure. Health Serv Res 40(6p1):1918–1930
19. Hibbard JH, Stockard J, Mahoney ER, Tusler M (2004) Development of the patient activation measure (pam): conceptualizing and measuring activation in patients and consumers. Health Serv Res 39(4p1):1005–1026
20. Holman H, Lorig K (2004) Patient self-management: a key to effectiveness and efficiency in care of chronic disease. Public Health Report 119(3):239–243
21. Karanasiou GS, Tripoliti EE, Papadopoulos TG, Kalatzis FG, Goletsis Y, Naka KK, Bechlioulis A, Errachid A, Fotiadis DI (2016) Predicting adherence of patients with hf through machine learning techniques. Healthcare Technol Lett 3(3):165–170
22. Kilic B, Kalaca S, Unal B, Phillimore P, Zaman S (2015) Health policy analysis for prevention and control of cardiovascular diseases and diabetes mellitus in turkey. Int J Public Health 60(1):47–53
23. Koesmahargyo V, Abbas A, Zhang L, Guan L, Feng S, Yadav V, Galatzer-Levy IR (2020) Accuracy of machine learning-based prediction of medication adherence in clinical research. Psychiatry Res 294:113558
24. Kosar C, Besen DB (2019) Adaptation of a patient activatıon measure (pam) into turkish: reliability and validity test. African Health Sci 19(1):1811–1820
25. Lindsay A, Hibbard JH, Boothroyd DB, Glaseroff A, Asch SM (2018) Patient activation changes as a potential signal for changes in health care costs: cohort study of us high-cost patients. J General Int Med 33(12):2106–2112
26. Ma Y, He H (2013) Imbalanced learning: foundations, algorithms, and applications
27. McAllister M, Dunn G, Payne K, Davies L, Todd C (2012) Patient empowerment: the need to consider it as a measurable patient-reported outcome for chronic conditions. BMC Health Serv Res 12(1):1–8
28. McBain H, Shipley M, Newman S (2015) The impact of self-monitoring in chronic illness on healthcare utilisation: a systematic review of reviews. BMC Health Serv Res 15(1):1–10
29. Murdock RJ, Kauwe SK, Wang AY-T, Sparks TD (2020) Is domain knowledge necessary for machine learning materials properties? Integrat Mater Manufact Innovation 9(3):221–227
30. Norman P, Bennett P (1996) Health locus of control. In Predicting health behaviour: research and practice with social cognition models, pages 62–94. Open University Press
31. Obermeyer Z, Emanuel EJ (2016) Predicting the future-big data, machine learning, and clinical medicine. New England J Med 375(13):1216
32. O'Malley D, Dewan AA, Ohman-Strickland PA, Gundersen DA, Miller SM, Hudson SV (2018) Determinants of patient activation in a community sample of breast and prostate cancer survivors. Psycho-oncology 27(1):132–140
33. Osborne RH, Batterham RW, Elsworth GR, Hawkins M, Buchbinder R (2013) The grounded psychometric development and initial validation of the health literacy questionnaire (hlq). BMC Public Health 13(1):1–17
34. Poitras M-E, Maltais M-E, Bestard-Denommé L, Stewart M, Fortin M (2018) What are the effective elements in patient-centered and multimorbidity care? a scoping review. BMC Health Serv Res 18(1):1–9
35. Protheroe J, Rowlands G, Bartlam B, Levin-Zamir D (2017) Health literacy, diabetes prevention, and self-management. J Diabetes Res 2017:1298315
36. Queenan C, Cameron K, Snell A, Smalley J, Joglekar N (2019) Patient heal thyself: reducing hospital readmissions with technology-enabled continuity of care and patient activation. Product Oper Manag 28(11):2841–2853
37. Rao I, Shaham A, Yavneh A, Kahana D, Ashlagi I, Brandeau ML, Yamin D (2020) Predicting and improving patient-level antibiotic adherence. Health Care Manag Sci 1–13
38. Remmers C, Hibbard J, Mosen DM, Wagenfield M, Hoye RE, Jones C (2009) Is patient activation associated with future health outcomes and healthcare utilization among patients with diabetes? J Ambulat Care Manag 32(4):320–327
39. Sakarya S, Kulak E, Gorcin Karaketir S, Dogan E, Akman M, Cifcili S, Gunes E, Ormeci L (2019) Factors associated with patient activation in a turkish population with diabetes and/or hypertension. Eur J Public Health 29(Supplement_4):ckz186–225
40. Schwarzer R, Fuchs R et al (1996) Self-efficacy and health behaviours. Predicting health behavior: Research and practice with social cognition models 163:196
41. Shively MJ, Gardetto NJ, Kodiath MF, Kelly A, Smith TL, Stepnowsky C, Maynard C, Larson CB (2013) Effect of patient activation on self-management in patients with heart failure. J Cardiovasc Nurs 28(1):20–34
42. Simmons LA, Wolever RQ, Bechard EM, Snyderman R (2014) Patient engagement as a risk factor in personalized health care: a systematic review of the literature on chronic disease. Genome Med 6(2):1–13
43. Sinha AP, Zhao H (2008) Incorporating domain knowledge into data mining classifiers: An application in indirect lending. Decis Support Syst 46(1):287–299
44. Smith SG, Curtis LM, Wardle J, von Wagner C, Wolf MS (2013) Skill set or mind set? associations between health literacy, patient activation and health. PloS one 8(9):e74373
45. Son Y-J, Kim H-G, Kim E-H, Choi S, Lee S-K (2010) Application of support vector machine for prediction of medication adherence in heart failure patients. Healthcare Inf Res 16(4):253–259
46. Tripoliti EE, Papadopoulos TG, Karanasiou GS, Naka KK, Fotiadis DI (2017) Heart failure: Diagnosis, severity estimation and prediction of adverse events through machine learning techniques. Comput Struct Biotechnol J 15:26–47
47. Wallert J, Gustafson E, Held C, Madison G, Norlund F, von Essen L, Olsson EMG (2018) 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. J Med Internet Res 20(10):e10754
48. Wang H, Zheng B, Yoon SW, Ko HS (2018) A support vector machine-based ensemble algorithm for breast cancer diagnosis. Eur J Oper Res 267(2):687–699
49. Weil AR (2016) The patient engagement imperative. Health Affairs 34(4)
50. Wilcox AB, Hripcsak G (2003) The role of domain knowledge in automating medical text report classification. J Am Med Inf Assoc 10(4):330–338

51. Wong ST, Peterson S, Black C (2011) Patient activation in primary healthcare: a comparison between healthier individuals and those with a chronic illness. Med Care 469–479
52. Zhou M, Fukuoka Y, Goldberg K, Vittinghoff E, Aswani A (2019) Applying machine learning to predict future adherence to physical activity programs. BMC Med Inf Decision Making 19(1):1–11

Authors and Affiliations

Onur Demiray[1] · Evrim D. Gunes[2] · Ercan Kulak[3] · Emrah Dogan[4] · Seyma Gorcin Karaketir[5] · Serap Cifcili[6] · Mehmet Akman[6] · Sibel Sakarya[7]

Onur Demiray
o.demiray21@imperial.ac.uk

Serap Cifcili
serapcifcili@gmail.com

Mehmet Akman
makman4@gmail.com

Sibel Sakarya
ssakarya@ku.edu.tr

1 Department of Computing, Imperial College London, London SW7 2AZ, UK

2 College of Administrative Sciences and Economics, Koç University, Rumeli Feneri Yolu, Sariyer-Istanbul, Turkey

3 Ministry of Health Caycuma District Health Directorate, Zonguldak, Turkey

4 Ministry of Health, Zonguldak Community Health Center, Zonguldak, Turkey

5 Department of Public Health, Istanbul University, Istanbul, Turkey

6 Department of Family Medicine, Marmara University School of Medicine, Istanbul, Turkey

7 MPH, MHPE, School of Medicine, Department of Public Health, Koç University, Rumeli Feneri Yolu, Sariyer-Istanbul, Turkey