Health Information Science and Systems

RESEARCH

Who determines United States Healthcare out-of-pocket costs? Factor ranking and selection using ensemble learning

Chengcheng Zhang[1*], Yujia Ding[1,2] and Qidi Peng[1,2]

Abstract

Purpose: Healthcare out-of-pocket (OOP) costs consist of the annual expenses paid by individuals or families that are not reimbursed by insurance. In the U.S, broadening healthcare disparities are caused by the rapid increase in OOP costs. With a precise forecast of the OOP costs, governments can improve the design of healthcare policies to better control the OOP costs. This study designs a purely data-driven ensemble learning procedure to achieve a collection of factors that best predict OOP costs.

Methods: We propose a voting ensemble learning procedure to rank and select factors of OOP costs based on the Medical Expenditure Panel Survey dataset. The method involves utilizing votes from the base learners *forward subset selection*, *backward subset selection*, *random forest*, and *LASSO*.

Results: The top-ranking factors selected by our proposed method are *insurance type*, *age*, *asthma*, *family size*, *race*, and *number of physician office visits*. The predictive models using these factors outperform the models that employ the factors commonly considered by the literature through improving the prediction error (test MSE of the OOP costs' log-odds) from 0.462 to 0.382.

Conclusion: Our results indicate a set of factors which best explain the OOP costs behavior based on a purely data-driven solution. These findings contribute to the discussions regarding demand-side needs for containing rapidly rising OOP costs. Instead of estimating the impact of a single factor on OOP costs, our proposed method allows for the selection of arbitrary-sized factors to best explain OOP costs.

Keywords: Out-of-pocket costs, Health insurance, Variable importance rankings, Ensemble learning

Mathematics Subject Classification: 62P10, 62P20, 62J12, 62J07

Introduction and literature review

Background and motivation: healthcare expenditures, out-of-pocket costs and factor importance analysis

Healthcare out-of-pocket (OOP) costs consist of the annual expenses paid by individuals or families that are not reimbursed by insurance, including deductibles, coinsurance, copayments and services. The growth in OOP costs in the United States is fast. For example, OOP costs rose from $539 per person in the mid-1990s to more than $1,125 in 2017 [1]. This rapid increase may raise economic concerns of households, as high individual and family OOP costs could cause a financial burden to patients [2]. For instance, compared with not having cancer, being actively treated for cancer increases the mean out-of-pocket costs by $1,170 [3]. To prevent this financial burden caused by OOP costs from broadening healthcare disparities, we need to deeply understand who causes high OOP costs. Various factors have been discovered by the healthcare literature, such as *type of health insurance coverage*, *health status*, *demographics* and *socioeconomic status* [4–13]. However, each of the above references discusses of only one or two factors of OOP costs and no one considers studying the above factors jointly

*Correspondence: chengcheng.zhang@cgu.edu
[1] Claremont Graduate University, Department of Economic Sciences, 150 E. 10th Street, California, Claremont, USA
Full list of author information is available at the end of the article

and ranks their importances. We feel necessary to fill this gap, as discovering the OOP factor importance rankings is crucial for several reasons. First, nowadays information search costs for investigators remain high, then discovering key determinants allows investigators to better predict respondents' healthcare spending behavior, and to allocate resources accordingly at no additional cost. Secondly, a focus on deriving information on the key determinants helps investigators to design surveys that collect core and precise information more efficiently. This efficiency will further enable investigators to derive information from a larger population. Finally, through understanding how the key determinants influence the OOP costs, governments can better adjust their strategies in designing healthcare policies, in order to better control the household's future healthcare OOP costs.

Ranking factor importance is a subject in data analysis. In a supervised learning problem with multiple factors, ranking factor importance involves creating a quantitative measurement to evaluate each factor's impact on the response variable and compare them. As mentioned previously, there is no lack of literature on studying how a *single factor* explains OOP costs. For example, Hwang et al. found a strong linear relationship between OOP costs and "the number of chronic conditions" [9]. Gwet and Machlin showed that the percentage of the population with high OOP costs decreases as "household income" goes up [5]. Although the disparities in OOP costs among different subgroups have been revealed, this gives no information on which factor is more important than the others when they are considered to jointly explain the OOP costs. So far as we know, there is virtually no study on the joint impact of all aforementioned determinants on OOP costs.

Driven by the above issues, the paper is devoted to studying the rankings of factors who determine the OOP costs in the United States, by using an efficient *self-designed voting ensemble learning procedure*. Our data-driven analysis not only discovered the top six factors who cause high OOP costs, but also provided several predictive models of the OOP costs based on them. These predictive models indicate clearly how each factor explains the OOP costs and help to forecast the future OOP costs.

Voting ensemble for factor ranking and selection

Ensemble learning is the process of combining multiple independent strategies to solve a particular machine learning problem. The method involves first running multiple alternative base learners to train the data, then picking the one with the best performance under some measurement [14]. Ensemble learning has higher flexibility than each of its base learners, with possibly the drawback of high computational complexity. However, thanks to the modest size of our raw dataset (39, 246 rows by 40 columns, see Sect. 2) it is no longer an issue in our setting. As one type of ensemble learnings, the *voting ensemble* is used to rank and select factors. In a voting ensemble, a factor gets one vote if it is chosen by a variable selection algorithm in this ensemble, then its importance is measured by the total vote count.

In this paper, we choose forward stepwise, backward stepwise subset selection approaches, and two variable ranking methods based on random forests as base learners, then use a self-designed voting ensemble to combine the results. *Our voting ensemble assumes that a factor gets one vote from a base learner if it is ranked among top five by that base learner.*

Now we briefly introduce the above base learners. Linear model subset selection involves comparing the goodness-of-fit of all possible linear models, each based on an identified subset of variables. Depending on how this subset is selected, this class of approaches includes best, forward stepwise and backward stepwise subset selection methods [15]. Best subset selection picks the subset with optimal goodness-of-fit out of all possible subsets of variables; forward stepwise subset selection adds one variable at each step to the best subset obtained in the previous step; backward stepwise subset selection removes one variable at each step from the best subset obtained in the previous step. Random forests is itself a powerful ensemble learning tool for variable importance rankings. Ranking variables using random forests involves applying a two-stage strategy that is based on a preliminary ranking of the explanatory variables using the random-forests-permutation-based score of importance [16]. Among all the above algorithms, no one uniformly dominates the others. All the procedures taken in our paper have been implemented in *R* and the source code is publicly shared on GitHub.[1]

The rest part of the paper proceeds as follows. Section 2 derives and describes the raw data. Section 3 is devoted to data preprocessing. Section 4 details the ensemble learning procedure and its results. Section 5 discusses of the outcome of the ensemble learning procedure when the OOP cost levels are defined differently. Finally, Sect. 6 concludes.

[1] https://github.com/health-care-cost-data-analysis/factor-ranking-and-selection.

Raw data extraction and description

Our raw dataset is extracted from the Medical Expenditure Panel Survey (MEPS) [17] in 2016 and 2017. The MEPS[2] datasets are derived from a national representative sample of the United States civilian noninstitutionalized population. They provide information on healthcare utilization and costs, types of health insurance coverage, health status, as well as a variety of socioeconomic and demographic characteristics.[3] The MEPS consists of three components: household, medical provider and insurance. We pick the one-year consolidated data from the household component as the raw dataset [18], which consists of 39, 246 observations (rows) and 40 variables (columns).

To analyze the effects of each determinant on OOP costs, we treat "out-of-pocket costs" (measured in dollars) in the raw dataset as the response variable. As we are interested in detecting the level change in OOP costs rather than the value change, we replace "out-of-pocket costs" with a two-level categorical variable: Level 0 - "OOP spending from $0 to $1,000"; Level 1 - "OOP spending more than $1,000". The rest 39 variables in the raw dataset are viewed as potential factors of the OOP cost level. Next, we assign all variables into five groups (see Table 1) and explain the rationales.

The group *healthcare costs* contains only the response variable *out-of-pocket*. It denotes the OOP cost level built based on the total amount of payment paid by individuals and families that provided in the MEPS. Taking this OOP cost level as response variable our goal becomes to determine whether individuals had spent over $1,000 *out-of-pocket* at the time of the survey in 2016 and 2017.

A number of variables in the group *demographics* have been demonstrated to yield a significant disparity in healthcare spending (e.g., *sex*, *race*, *region* and *age*). Females spend considerably more OOP than males. In particular, females aged between 19 and 44 spent an average of 65% OOP more than males largely due to maternity care costs [19]. Additionally, numerous studies have focused on the effects of racial disparities on OOP costs and access to care. One study concludes that Blacks and Hispanics receive sufficiently different care at a higher cost level than Whites [20]. Furthermore, the causal relationship between immigrants and OOP costs has been assessed [21]. Foreign-born individuals have fewer regular sources of care and, as a result, incurred lower costs than U.S.-born individuals. Moreover, there are further differences based on ethnicity, region and English proficiency among those not born in the United States. Among the above demographics factors, no existing evidence was obtained to suggest that one is more influential than the others.

The group *socioeconomics* mainly consists of two classes of variables: *income-related* and *employment-related* variables. *Income* is the most widely used measure of economic resources in United States health research. Healthcare is *a normal good* because the estimated income elasticity of the demand for healthcare ranges from 0.0 to 0.2 [22]. This shows that when income grows, the demand for healthcare services also rises. *Employment* also impacts OOP costs because when employment status changes, the type of insurance coverage may also change accordingly. Most Americans under age 65 rely on health insurance offered by their workplace as employer-sponsored health insurance (ESI) is the majority of private health insurance in the United States. Employers who enroll in ESI contribute to the cost of coverage for employees.

Type of health insurance coverage in the group *health insurance* has been shown by numerous studies to play a key role in explaining OOP costs [7, 23]. In our raw dataset, *types of health insurance coverage* consist of uninsured, private and public insurances. Uninsured respondents are required to pay all OOP themselves. Private insurance refers to plans provided by private companies that can be purchased by individual consumers or offered by employers. Premiums, deductibles and their OOP amounts vary by plans. Public insurance consists of Medicaid, TRICARE and so on. Medicaid is the primary public health coverage provided by the government for low-income individuals or families. Medicaid beneficiaries pay no, or a very low premium, and OOP expenses are based on income. Differences in OOP costs are often yielded by different *types of health insurance* plans because coverage may change patients' choices. For example, when a moral hazard occurs, individuals have an incentive to make more doctor visits if they have a low OOP health insurance plan [24]. To specify individuals' healthcare behavior, we have included *number of physician office visits* to catch the event of doctor visits in the group *health status* below.

Lower *health status* is often associated with higher OOP costs. Among all the health status information provided in MEPS, we pick self-reported health and mental status, six functional limitations and ten chronic health conditions to form our raw dataset. Functional limitations, which can be described as an impairment in an individual's ability to function, can be linked to chronic conditions. Chronic conditions often have long-term effect and require ongoing medical care, such as diabetes and asthma. All ten chronic conditions identified in

[2] http://www.meps.ahrq.gov/.

[3] http://meps.ahrq.gov/mepsweb/data_stats/download_data/pufs/h129/h129doc.shtml.

Table 1 Grouped variables

Group		Variable
Healthcare costs		**Out-of-pocket costs***
Demographic		Age
		Sex
		Race
		Region
		Family size
		Primary language not English
		English proficiency
		Marital status
		Born in the U.S.
		Years in the U.S.
		Year
Socioeconomic		Family income
		Individual's wage income
		Hourly wage level
		Employment status
		Self-employment status
		Occupation groups
		Purchased food stamps
Health status	Chronic condition	High blood pressure
		Coronary heart disease
		Stroke
		Bronchitis
		High cholesterol
		Cancer
		Diabetes
		Asthma
		Arthritis
		Joint pain
	Functional limitation	Serious hearing difficulties
		Serious seeing difficulties
		Serious cognitive difficulties
		Cognitive limitation
		Physical functioning limitation
		Work/Housework/School limitation (Any limitation)
		Used assistive devices
	Self-reported health status	Perceived health status
		Perceived mental health status
		Number of physician office visits
Health insurance		Type of health insurance coverage

* stands for the response variable; Only *Age, Family size, Family income, Individual's wage income* are numerical factors, the rest are categorical factors

the literature are associated with OOP costs [11], causing individuals with chronic conditions to exhibit a greater demand for healthcare. Moreover, Hwang et al. found that the OOP costs increase as the number of chronic conditions increases [9]. Recently it is shown that chronic conditions are the major reason for 37% of office-based physician visits [25]. As the demand for healthcare service increases, the *number of physician office visits* increases. As a result, the probability of spending more on OOP costs increases. Finally, we conclude that people with a lower *health status* require more care which directly impacts their OOP costs.

Even though there is ample literature that reveals the impact of the above factors, it is still uncertain about which ones are dominating factors. Therefore, we need to establish a measurement of variable importance to compare these factors' effects in the OOP costs. As mentioned in Sect. 1.2, an ensemble learning tool can help to acquire this goal. In Sects. 3–4 below, we show how this ensemble learning is processed and analyze its output at each step.

Data preprocessing

The raw dataset (see Table 1) is not ready yet to be put through the ensemble learning process for the following reasons: (1) MEPS survey is open for all age groups, but here we focus our study only on the working age population. Therefore, respondents aged below 18 or above 64 can be excluded. (2) Missing value occurs when a respondent misses or refuses to answer a question. In the survey, these values may be marked as *inapplicable, don't know (DK)*, etc. This issue makes some variables consist of an unnecessarily large number of blank information, such that the dataset faces the "sparse data issue". (3) Due to the large number of variables in the raw dataset, there may exist strong dependencies among them, which yields inconsistent and misleading variable selection outputs.

To overcome the above issues, it is necessary to perform data preprocessing before running the variable ranking algorithms. Our data preprocessing process consists of two stages: data engineering (see Sect. 3.1) and data cleaning (see Sect. 3.2).

Data engineering

In this section, we filter the survey records and rebuild some variables' categories. First, we focus our attention on studying the behavior of adults of working age because otherwise, the discrepancy in health status is too large. Therefore, we extract all records with ages ranging from 18 to 64. Secondly, as MEPS presents a "sparse data issue": a large number of low-frequency categories are observed, so that the predictive model can barely capture them. To overcome this issue, we merge categories which share similar patterns, i.e., we regard the responses *don't know*, *not ascertained*, *refused* and *inapplicable* as missing values and mark them all as "-1". MEPS describes *marital status* as *married*, *widowed*, *divorced*, *separated* and *never married*. Thus, we merge the latter four categories into a single one "*unmarried*". For *employment status*, we label both *job to return* and *job during* as "*job to return/job during*". Finally, MEPS also asks respondents for *hourly wage*, resulting in a variable with mixed-type (numerical and categorical) data. Because there is no simple machine learning approach in the literature that trains this data type, we transform it into a categorical variable by dividing it into four categories: individuals who earn "\$0 *to* \$40", "\$40.01 *to* \$85", "*more than* \$85", and "-1" (missing values). The new dataset is then ready for being analyzed, which is downloadable in GitHub.[4]

Data cleaning: dependency detection and variables removal

In this step, we identify groups of dependent variables using the correlation detection method in [26], then remove all the redundant variables. Correlation detection is considered an important step of data cleaning to reduce collinearity, when the data set contains highly correlated variables. In the literature, many correlation measurements have been studied, but each assumes that the paired variables follow a specific type and that improper usage will output a nonsensical result. To properly deal with a data set that involves nominal, ordinal, and numerical variables, we use the correlation coefficient ϕ_K that was first introduced in [26]. The ϕ_K[5], valued between 0 (uncorrelated) and 1 (totally correlated), works consistently between pairs of numerical, nominal, and ordinal variables and captures both linear and nonlinear dependencies. It is obtained by applying the χ^2 contingency test using Pearson's χ^2 test statistic and the statistically dependent frequency estimates, then interpreting the χ^2 value as coming from a bivariate normal distribution; the corresponding correlation parameter is the ϕ_K.

In practice, the levels of correlation and significance should always be studied together, because a large correlation may be statistically insignificant and vice versa. Therefore, we recognize pairs that satisfy "$\phi_K \geq 0.9$ and has a higher-than-median significance level" as having a high correlation. The high correlated pairs are summarized in Table 2 in the sense that pairs of the variables in the same subgroup are shown to satisfy the criteria. Since variables in each subgroup are of equal importance, it suffices to arbitrarily pick one representative (followed by * in Table 2) from each subgroup and drop the others.

In Table 2, the factors in each subgroup are strongly dependent of each other. Next, we explain the rationales of this correlation detection result:

Subgroup I is a five-variable subset of the group "Functional limitations and self-reported health status". These five variables have a strong relationship with each other. Indeed, it is reasonable for people who used assistive devices to perceive that they are in insufficient physical or mental health status. Individuals who self-reported poor

[4] https://github.com/health-care-cost-data-analysis/factor-ranking-and-selection/blob/master/MEPS_data.csv.

[5] The analyzer is available as a Python library through the PyPi server: https://phik.readthedocs.io/en/latest/.

Table 2 Correlation detection

Group	Dependent variables	Label
Functional limitations and self-reported health status	Any limitation*	Subgroup I
	Used assistive devices	
	Physical functioning limitation	
	Perceived health status	
	Perceived mental health status	
	Serious cognitive difficulties*	Subgroup II
	Serious hearing difficulty	
	Serious seeing difficulty	
Chronic condition	Diabetes*	Subgroup III
	Cancer	
	Stroke	
	Arthritis	
	Coronary heart disease	
	High cholesterol	
	High blood pressure	
	Joint pain*	Subgroup IV
	Bronchitis	

* stands for the variable selected to represent its subgroup. Thirteen variables have been removed after this stage

mental and physical health were likely to have physical functioning limitations as well. All of these restrictions affect work, housework, and/or school which are marked as *any limitation*. In conclusion, we let *any limitation* be the representative variable and drop the rest three.

Subgroup II is another subset of the group "Functional limitations and self-reported health status". The literature has shown that hearing and seeing difficulties heavily influence cognitive difficulties. For instance, one study found a strong relationship between cognitive function and hearing loss [27]. In addition, both seeing and hearing abilities are related to age since functions start to decline once one reaches a certain age [28]. Therefore, the three variables in this subgroup are highly dependent on each other. Consequently, we select *serious cognitive difficulties* to be the representative variable of Subgroup II.

Subgroups III & IV both belong to the group "Chronic condition". They consist of specific types of diseases. Some diseases are considered important factors since they straightforwardly lead to high OOP costs. The diseases themselves may depend on each other. First, some diseases create multiple complications. For example, those with *diabetes* are more likely to have *heart disease* or *stroke* than those without *diabetes* [29]. Next, certain behavioral habits may create multiple chronic diseases. For example, obese individuals have an increased risk of developing a number of chronic diseases, such as diabetes, asthma and cancer [30]. Another example of a behavioral habit is smoking. It is well documented that smoking is a risk factor for stroke, asthma, heart disease and several cancers. In our dataset, seven chronic conditions inter-correlate in Subgroup III. Moreover, correlation detection shows joint pain is dependent on bronchitis in Subgroup IV. To resolve the inter-dependent issue, we hold *diabetes* from the seven conditions in Subgroup III and *joint pain* in Subgroup IV.

In addition to the dependency threshold of 0.9, we also perform the correlation detection at the level of 0.8, 0.7 and 0.6. As we decrease the level of dependency, more variables become dependent with others. For example, when the correlation coefficient $\phi_K \geq 0.8$, the dependency causes six more variables to be marked for deletion, compared to the 0.9 level. When the level is at 0.6, *joint pain*, *asthma* and the other four variables are manually picked from the correlation detection output, and 25 variables need to be removed from the dataset. Variables such as *type of insurance coverage*, *joint pain*, *asthma* and *primary language* remain at the levels 0.6 to 0.9. Because our objective is to involve as many variables as possible in order to reveal their impact on OOP costs, we choose a high dependency level of 0.9. Based on this correlation detection result and our manual picks, we preserve the variables: *any limitation, serious cognitive difficulties, diabetes* and *joint pain* and drop the rest in Table 2. Consequently, thirteen variables are removed from the raw dataset.

Ensemble learning procedure

This section introduces the main fruit of our framework. We perform an ensemble learning for variable selection on the dataset obtained from Sect. 3. This ensemble learning is designed as follows:

Step 1: Response variable transformation. We run a logistic regression to fit the OOP cost levels using all the 26 variables as predictors. This step aims at transforming the OOP cost levels into the log-odds. Let $\mathbb{P}(A) = p \in (0, 1)$ be the probability that the event A occurs, the log-odds of the event A is defined to be $\log\left(\frac{p}{1-p}\right)$, which is valued in $(-\infty, +\infty)$. In this step, we replace the OOP cost levels with its log-odds in the dataset. Such "categorical to numerical data" transformation will enable us to apply the three subset selection methods (see *regsubsets()* from the package *leaps* in *R*) in the next step. Note that this efficient method *regsubsets()* employs different model selection criteria such as C_p, AIC, BIC, which differ only in how models of different sizes are compared. Therefore, the results do not depend on the choice of cost-complexity trade-off. More detail will be given in Sect. 4.1.

Step 2: Variable selection. We respectively perform the best subset selection, random forests and LASSO to pick the best subset variables out of the 26 variables, using the log-odds of OOP cost levels as the response variable.

Step 3: Variable votes ranking. Note that forward stepwise subset selection, backward stepwise subset selection and random forests are able to provide variable importance rankings. For each of the 26 variables, count one vote if its ranking is among the top five in one of the following four measurements: (1) forward stepwise subset selection, (2) backward stepwise subset selection, (3) mean decrease in accuracy (output from random forests), and (4) mean decrease in node purity (output from random forests). Output the total votes count of each variable. The variable votes rankings are obtained through sorting the above votes counts. We will see using this "top five" criterion allows us to select six leading factors (votes count ≥ 2) of the OOP cost levels, which is a proper size for both OOP costs behavioral analysis and predictive model implementation.

Step 4: Model validation. We first use 5-fold cross-validation approach to estimate the test mean squared errors (MSEs) of the best, forward stepwise and backward stepwise subset selections. Each is performed with its best subset variables. We then compare these test MSEs to the ones output by the ridge regression, the LASSO, the random forests, the linear model with the voted variables and the linear model with the literature supported variables.

Sections 4.1–4.7 below will be in charge of the above Steps 1–4 stepwisely.

Response variable transformation: additive logistic regression

In this stage, we use an overfitting additive logistic regression to transform the categorical OOP cost levels to numerical log-odds. Recall that the response variable "OOP costs" consists of two levels, where Level 0 denotes the event "OOP spending not more than \$1,000 per year" and Level 1 denotes the event "OOP spending over \$1,000 per year". Mathematically, denote by y the level of OOP costs; let X be the list of predictors and $x = (x_1, \ldots, x_{26})$ be the 26 candidate variables in our dataset; let $\mathbb{P}(y = 1|X = x)$ denote the probability of $y = 1$ given $X = x$. This is the chance that an individual in status $(x_1, \ldots, x_{26})$ (sex, age, race, etc.) spends OOP over \$1,000 per year.

The reason why we choose the classifier "additive logistic regression" to fit y, is that y is a binary-label variable and most of the factors in x are categorical variables. The method involves solving β from the following equation:

$$\begin{cases} \mathbb{P}(y = 1|X = x) = \frac{1}{1+e^{-z}}; \\ z = \beta_0 + \beta \cdot x. \end{cases} \tag{1}$$

where z is the log-odds of the event $y = 1$ given $X = (x_1, \ldots, x_{26})$. The intercept β_0 describes the basic level of the probability for the event $y = 1$ given $X = (x_1, \ldots, x_{26})$; $\beta = (\beta_1, \ldots, \beta_{26})$ are the coefficients or slopes, where β_k $(k \in \{1, \ldots, 26\})$ measures the effect of the kth factor x_k on the OOP cost levels y; $\beta \cdot x = \beta_1 x_1 + \ldots + \beta_{26} x_{26}$ is the part of z explained by the effects x. If x_k $(k \in \{1, \ldots, 26\})$ is a numerical variable, $\beta_k x_k$ then denotes the conventional product. If x_k is categorical with m $(m \geq 2)$ categories $c_1, \ldots, c_m$, then in an additive model the notation $\beta_k x_k$ is viewed as

$$\beta_k x_k = \begin{cases} \beta_k^{(1)} & \text{if } x_k = c_1; \\ \ldots & \\ \beta_k^{(m-1)} & \text{if } x_k = c_{m-1}; \\ 0 & \text{if } x_k = c_m. \end{cases} \tag{2}$$

Hence, the additive logistic regression works by adding some "weight" to z for each category.

Suggested and popularized by Hastie and Tibshirani [31], the additive logistic regression has several advantages: First, it is free of distributional assumptions, i.e., the variables' probability distributions need not be known. Next, although not considered in our framework, it can easily handle interaction effects between factors in a flexible way. In addition, as a generalized linear model, it is characterized by a manageable number of coefficients that can be intuitively interpreted. Finally, it can be easily implemented by the method *glm*() in *R*.

Note that performing additive logistic regression over the 26 variables will overfit the OOP cost levels. This achieves our goal of transforming the response variable to real numbers since the training error is minimized through overfitting.

Variable selection: best subset selection

In the raw dataset, replacing the OOP cost levels with the log-odds obtained in Sect. 4.1, we are ready to apply the three base learners for variable selection: best subset selection, random forests and LASSO.

Best subset variable selection aims at selecting a subset out of 26 variables (containing a total number of 52 categories) that best explains the OOP costs under the linear model. By performing the best subset selection method using the method *regsubsets*() in *R*, we obtain the 1-factor to 52-factor best subsets, and present the comparison

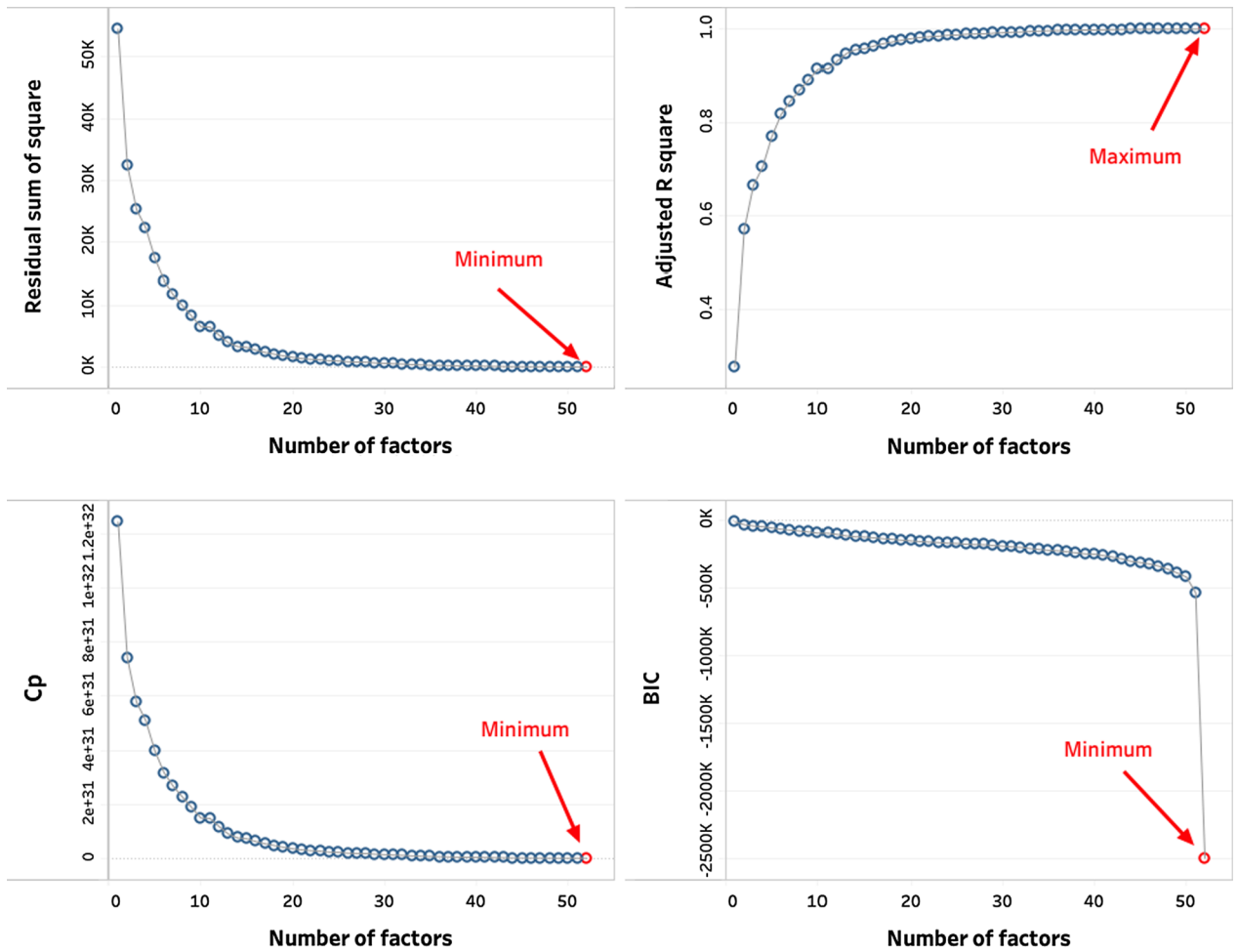

Fig. 1 Best subset's measurements of goodness-of-fit against "number of factors". The point in red corresponds to the optimized measurement value. Unlike the other three measurements, the BIC curve observes a dramatic decrease of 80% when the last factor is included. This presents an overfitting phenomenon

results of the four measurements of goodness-of-fit (RSS, adjusted R^2, C_p and BIC) in Fig. 1 below.

In Fig. 1, C_p, BIC, and adjusted R^2 all suggest the linear regression containing over 50 factors to be the best fitting model. However, this over-50-factor linear model suffers from an overfitting issue because after a sharp decrease from 1-factor to about 15-factor, the model's fitting errors start to be flat. In other words, the linear model starts to fit noise. Overfitting issue often yields the high-variance estimation. In our case, it entails that the number of factors is obsessively large so that the linear model loses the ability to describe the underlying relationship between healthcare determinants and OOP costs. To overcome this issue, we will try to pick a subset that lowers the test MSE.

Figure 2 represents the comparison results of the test MSEs estimated by using the 5-fold cross-validation approach. Although the lowest test MSE is obtained by fitting with 44 factors, we recommend the 29-factor subset as the best pick based on the following analysis and considerations.

On one hand, from the top chart in Fig. 2 we observe that the test MSE corresponding to the 29-factor subset is not significantly higher than its global minimum corresponding to the 44-factor subset. On the other hand, the incremental test MSE curve in the bottom chart is negative-valued until the number of factors equals 29. This indicates no overfitting issue occurs if one selects up to 29 factors. Moreover, we would avoid the increase of model complexity from 29-factor to 44-factor. This is because a complex healthcare model risks being overly costly. It refers to the amount of time spent on designing surveys, as well as taking them. It is also a potential problem for respondents to understand the surveys due

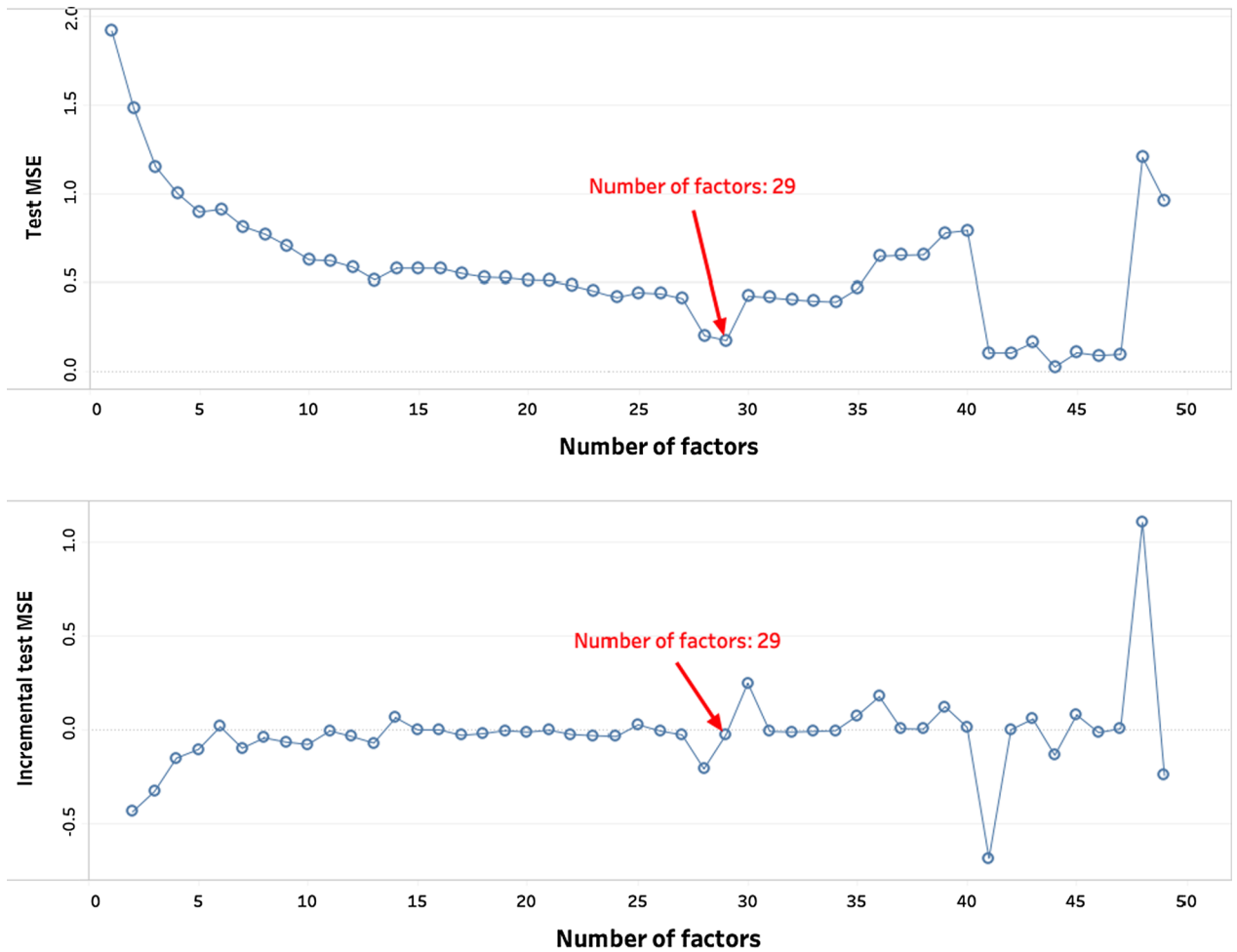

Fig. 2 Optimal test MSE. Top: test MSE against number of factors. Bottom: the magnitude of incremental test MSE against number of factors. The red arrow highlights the subset that contributes significant changes to the test MSE and that balances the estimation bias and variance

to an excess of questions. Given the above three facts, we believe that the 29-factor subset is the best pick.

Next, we explain why the 29-factor subset happens to be a good choice from its variables' features. This subset consists of 29 categories out of a total number of 52, coming from 19 variables out of a total number of 26. The red arrow in the figure indicates that the 29-factor subset yields a test MSE score of 0.176. The test MSE curve also observes its dramatic decline when the number of factors goes from 27 to 29: it drops almost 57.3% from 0.4119 to 0.176. What happens here is: From the 27-factor to 28-factor model, the category "*born in the U.S.: No*" is removed. Meanwhile, "*self-employment status: No*" and "*any limitation: Yes*" are added. From the 28-factor to 29-factor model, the numerical variable "*number of physician office visits*" is added. Note that the "*number of physician office visits*" represents individuals' levels of healthcare utilization. It also reflects individuals' health status. Hence, it reasonably impacts OOP costs. From Fig. 2 we see by adding this variable, the test MSE almost reaches its smallest value.

Table 3 displays our recommended best subset 29 factors (19 variables) $x = (x_1, \ldots, x_{19})$) and their coefficients $(\beta = (\beta_1, \ldots, \beta_{19}))$ in the logistic regression model:

$$\log\left(\frac{\mathbb{P}(y=1|X=x)}{1-\mathbb{P}(y=1|X=x)}\right) = \beta_0 + \beta \cdot x,$$

where β_0, $\beta \cdot x$ are defined as in (1). For $k \in \{1, \ldots, 19\}$, β_k measures the expected change in the log-odds of "OOP spending over $1,000" by a unit increase in the

Table 3 Factors selected by the best subset selection method

Variable	Factor	Coefficient
(Intercept)	**Numerical**	**(5.5985)**
Type of insurance coverage	Public health insurance	(1.3386)
	Uninsured & Private health insurance	—
Age	Numerical	0.0150
Asthma	No	3.0346
	Yes	3.1408
	Inapplicable, not ascertained, DK, Refused	—
Family size	Numerical	(0.1940)
Race	Black	(0.2781)
	White	0.5092
	Multiple races	—
	American India/Alaska native	—
	Asian/Native Hawaiian/PACFC ISL	—
Number of physician office visits	Numerical	0.1140
Family income	Numerical	3.2786×10^{-6}
Primary language not English	Spanish	(0.5704)
	Another language & Inapplicable	—
Sex	Male	(0.4506)
	Female	—
Joint pain	No	(0.3628)
	Yes & Inapplicable, not ascertained, DK, Refused	—
Purchased food stamp	No	0.6327
	Yes & Inapplicable, not ascertained, DK, Refused	—
Occupation groups	Farming, fishing and forestry	0.0499
	Management, business and financial	0.1821
	Military specific occupations	(0.6971)
	Sales and related occupations	0.0633
	Professional and related occupations	—
	Office and administrative support	—
	Production, transportation, matrl moving	—
	Service occupations	—
	Construction, extraction, maintenance	—
	Inapplicable, not ascertained, Unclassifiable	—
Diabetes	No	(0.3414)
	Yes	0.3178
	Inapplicable, not ascertained, DK, Refused	—
Any limitation	Yes	0.3443
	No & Inapplicable, DK, Refused	—
Marital status	Refused, DK	(8.3354)
	Widowed, Divorced, Separated, Never married	(0.1776)
	Married	—
Self-employment status	No	(0.0500)
	Yes	0.2052
	Inapplicable	—
Region	Northeast	(0.3176)
	West	(0.1647)
	South	(0.1464)
	Midwest & Inapplicable	—
Serious cognitive difficulty	No	(0.0566)
	Yes & Inapplicable, not ascertained, DK, Refused	—
Cognitive limitation	No	0.2595
	Yes & Inapplicable, not ascertained, DK, Refused	—

— denotes 0; (•) denotes a negative value

corresponding factor x_k. A positive coefficient indicates that the corresponding feature increases the chance of "OOP spending over $1,000" and vice versa.

In the following, we explain how the above selection results agree with the existing literature. Previous literature demonstrates that OOP costs are greatly impacted by health insurance coverage status, utilization and health status. However, as shown in Figs. 1 and 2, the 44-factor subset leads to the best bias and variance result, which means in the real world healthcare costs are affected in a more complicated way by a larger number of factors. This is true since in addition to the above three factors, individual healthcare behavior is also involved in explaining the OOP costs. First, healthcare costs are incurred if an individual visits a doctor or undergoes a medical test. Next, individuals with greater healthcare needs are more likely to purchase health insurance based on the idea of adverse selection. Finally, the more that healthcare services are used by individuals, the higher incidence of OOP costs depends on insurance plans. It is also important to take copayments and deductibles into account. However, such information was not provided by MEPS for 2016 or 2017. The following paragraphs are devoted to reviewing the literature on explaining how the factors insurance coverage, utilization and health status impact an individual's OOP costs. These factors are consistent with our data-driven solution in Table 3. We also make links between the new findings in Table 3 and OOP costs.

Type of insurance coverage directly impacts OOP costs. One reason is: a *health insurance* plan usually sets an upper limit on OOP costs. However, MEPS does not provide specific information on the upper limit amount of the insurance plan. Insured individuals only pay a portion of the bill based on their *health insurance* plans, whereas uninsured respondents pay everything OOP. According to previous studies, individuals with *health insurance* significantly lower their risk of incurring catastrophic expenditures [23]. Our regression result (See Table 3) shows that choosing public health insurance coverage makes the log-odds of "OOP spending over $1 ,000" decrease by 1.3386. This coefficient indicates that the probability of "spending over $1,000" for individuals with public insurance is approximately 21.89 percentage points[6] lower than private insurance beneficiaries. Because individuals with public health insurance lower the chance of spending over $1,000, we conclude that having public health insurance coverage includes some protection against increasing OOP costs.

As shown in Table 3, the coefficient of the numerical variable *family income* is 3.2786×10^{-6}. This means when *family income* increases by $10,000, the probability of "OOP spending over $1,000" increases by around 0.435 percentage points. Our study has shed light on the impact of *family income* on OOP costs which previous literature has shown. In 2015, 52% of poor families spent less than $100 OOP, and 4% spent over $2,500. By comparison, 11% of high-income families spent less than $100 OOP, and 22% spent over $2,500 [5]. The disparities in healthcare among *income* levels have been well documented [32]. Poverty levels are usually used to determine eligibility for certain medical programs and benefits. Many states require 100% of the federal poverty level (FPL) as the *income* limit for qualifying for adults Medicaid. As mentioned, those who are eligible for Medicaid, pay no, or a very low OOP expense. Our findings are aligned with reports in the literature that individuals with lower *family income* pay lower OOP.

The factors *age*, *sex*, *race*, *primary language* and *region* all belong to the demographics group. *Age* is a time index. OOP costs are heavily time-dependent, since a number of leading factors such as an individual's *health status*, *income* and *family size* may change with time. The coefficient of *age* (0.0150) entails that every year the probability of "OOP spending over $1,000" increases around 2.62 percentage points. *Sex* is an important categorical variable that captures an individual's behavior in healthcare. Table 3 shows that the probability of "OOP spending over $1,000" for males is approximately 7.82 percentage points less than for females (with the log-odds of −0.4506 for male). This indicates that females are likely to spend more OOP than males. The significant differences in healthcare costs across *age* and *sex* groups because different *age* and *sex* groups may face different health issues [19]. On one hand, both males and females take various healthcare services as they age. On the other hand, females spend significantly more than males largely due to maternity care [19]. Our data-driven solutions in *age* and *sex* are then in line with the existing literature.

The factor *race* affects a large group of other factors such as *primary language not English, place of birth, English proficiency*. The latter three factors may affect the individual's ability to understand insurance benefits or access to healthcare. Based on our analysis, all race-related variables are considered important in impacting OOP costs. We first discuss the effects of the *race* itself. According to Table 3, Blacks (−0.2781) and Whites (0.5092) have opposite signs of coefficients. The probability of "OOP spending over $1,000" for Black is around 26 percentage points lower than White. Racial inequalities are found in many sectors of American life. Minority populations continue to face an imbalance in healthcare

[6] One can calculate the increment of probability through $1 - \frac{1}{1+e^{\Delta l} p_0/(1-p_0)} - p_0$, given the corresponding base probability p_0 and the increment of log-odds Δl.

and health. Racial disparities have also been documented by previous studies that show *race* to be one of the most important factors of OOP costs [20, 33].

Primary language not English is a factor that relates to *race*. Spanish speakers have a negative impact (−0.5704) on OOP costs. For those whose primary language is not English, individuals whose *primary language* is Spanish have about 9.95 percentage points lower on the probability of spending over $1,000 compared to other language speakers. Hence, our study reveals the impact of a *primary language* spoken at home aside from English. There is an inadequate amount of literature that studies the relationship between the *primary language* spoken at home and OOP costs. However, there exists an abundance of literature that reveals the lower rates of healthcare services among non-English speakers in the United States [34–36]. The observed disparities may be attributable to the enduring effects of language, such as income and health insurance status. However, it is more likely that these factors are not the most important to healthcare utilization but play a secondary role. Non-native speakers may have problems earning a living that matches their abilities and intellect. They also may have trouble finding a source of care. Communication difficulties among those whose primary language is not English put them at risk for receiving eligible healthcare services compared to native English household speakers [37]. Hispanics whose *primary language* is not English were less likely to receive all eligible healthcare services. Regulating income and adjusting sources of care by itself miscalculates the impact of not speaking English at home regarding healthcare utilization, so it will not solve the problem for people whose *primary language* is not English regarding OOP costs.

According to Table 3, respondents living in different *regions* such as the northeast, west and south all have negative coefficients. This corresponds to a lower chance of "OOP spending over $1,000" than not living in the above three regions. One possible reason is that every state of the United States has its own health insurance policy, which has affected health utilization and thus OOP costs. For example, California adopted Medicaid expansion through the Affordable Care Act (ACA) in 2014, allowing adults with incomes up to 138% of the poverty level to be eligible for coverage. However, states without Medicaid expansion only cover up to 100% of the FPL. Another possible reason is that some healthcare services have wide price ranges across the United States [38]. For example, the price for knee replacement incurred the greatest discrepancies - from an average of $19,934 in Iowa to an average of $61,750 in New York. It may cause patients who live in states with high-expense healthcare to put off healthcare-related appointments due to the unaffordable costs. As the demand for healthcare service in high-cost states goes down, the chance of paying more OOP decreases. Based on our results and previous studies, we see that *region* is one of the most important factors that impact people's access to, cost of, and quality of care; as a result, it heavily affects OOP costs.

As shown in Table 3, the best subset includes all the representative factors of the groups *chronic condition* and *functional limitations and self-reported health status* listed in Table 2. This signifies all the representative factors that have been selected by correlation detection in Table 2 strongly influence OOP costs. This finding has been supported by the literature [7, 8]. High OOP costs are also associated with painful health conditions, such as *joint pain*. For those who experience long-lasting pain, complementary approaches often help to manage painful conditions [39]. Millions of adults use complementary health approaches. This leads an individual's annual OOP costs to range from $568 to $895 [40]. Non-elderly adults who reported receiving treatment found medications and physician office visits prohibitively expensive [12]. Another common chronic disease is *diabetes*. Over 10% of the United States population suffers from *diabetes*. The financial burden related to *diabetes*, due to high OOP treatment costs, is made evident in the studies [23, 41].

The impact of all the remaining health status factors on OOP costs has been revealed by best subset selection method. For example, as *number of physician office visits* increase by 1 time, the probability of spending over $1,000 increase by around 2.19 percentage points. Another new finding is that, unlike *joint pain* and *diabetes*, the coefficients of with or without *asthma* are both positive. In other words, as long as respondents answered "Yes" or "No" to this question in the MEPS, the chance of spending over $1,000 increased. On one hand, it has been revealed that individuals with *asthma* tend to spend more OOP [12, 42]. Individuals with asthma are around 10.62 percentage points higher in the probability of spending over $1,000 compared to those without. On the other hand, it is also possible that OOP costs increase for individuals who answered "No" because they might schedule regular checkups or other medical treatments to verify they do not have *asthma*. Both the medical checkups and treatments increase the probability of spending more OOP.

Although a number of chronic conditions are proved to have a key impact on OOP costs, there is virtually no literature assessing which condition is more weighted than the others. In contrast, our data-driven solution enabled such comparison using the correlation test and coefficients: (1) From Table 2, the factors in each of Subgroups III and IV have almost equal weights on explaining OOP costs. (2) Given the coefficients in absolute

value in Table 3, we conclude that *asthma* is much more weighted (around 3) than *joint pain* (0.3628) and *diabetes* (around 0.33), while *joint pain* is slightly more weighted than *diabetes*.

Variable selection: random forests and LASSO

At this stage, we perform variable selection with two approaches: random forests and LASSO. Random forests are a class of attractive non-parametric, model-free and well-fitting statistical approaches. They are an ensemble learning method based on performing multiple decision trees. In this paper, we employ recently developed variable selection methods based on random forests – "Variable Selection using Random Forests" (VSURF) to select three subset variables: "thresholding subset", "interpretation subset" and "prediction subset". The VSURF method is first introduced by Genuer et al. [43] and has an implementation in *R* (see the method VSURF() in *R*).

The VSURF method consists of a two-step procedure. The first step is the preliminary elimination and ranking. We rank the variables by sorting the variable importance (VI) (increase in mean of the MSE of a tree) in descending order. Next, we eliminate the variables of small importance. More precisely, we drop the variables whose standard deviations of VI are less than some threshold value. This threshold value equals the minimum prediction value given by a classification and regression tree (CART) model. The next step is selecting interpretation and prediction subsets. For interpretation, we construct the nested collection of random forests models involving the first k variables, for $k = 1, \ldots, m$. We then select the variables yielding the smallest out-of-bag (OOB) error. This leads to considering m_0 $(< m)$ variables. For prediction, the variables are picked from the interpretation subset. We construct an ascending sequence of random forests models, by invoking and testing the variables in a stepwise way. The variables of the last model were selected. The test is performed as follows: a variable is added only if the error decrease is larger than the threshold given by

$$\frac{1}{m - m_0} \sum_{k=m_0}^{m-1} \left| OOB(k+1) - OOB(k) \right|,$$

where $OOB(k)$ is the OOB error built using the k most important variables.

The VSURF results are provided in Fig. 3, which displays different VI scores against the number of factors. We see that the thresholding, interpretation and prediction subsets contain 44 (19 variables), 39 (19 variables), 22 (15 variables) factors respectively. In Table 4 we observe that the selected variables are quite consistent with the ones selected by the best subset selection results in Table 3.

LASSO is a regularized linear model introduced and popularized by Tibshirani [44] which tries to remedy the linear regression's overfitting issue. The LASSO coefficients may be shrunk to 0. This feature makes it a variable selection approach. Nowadays, LASSO is widely used to reduce the overfitting issue and to improve prediction performance. In our analysis performing LASSO results in the selection of 25 variables. These 25 variables together with the ones picked by VSURF are listed in Table 4 below.

As the VSURF output, thresholding and interpretation are aligned with each other. Compared to the best subset selection result in Table 3 and the prediction subset by the random forests in Table 4, the LASSO has chosen almost all the 26 variables. Therefore, it turns out to be more overfitting than the other two approaches. These selected subsets will be used as benchmark models in the model validation step (see Sect. 4.7).

Variable importance ranking: forward and backward stepwise subset selections

The forward and backward stepwise subset selection methods also help to select the subset of variables out of 26 (52 categories) for the final model. In our case, these two methods provide the same best subset of variables as best subset selection, serving as a strong proof that the subset selection approaches using the logistic regression are consistent (see Table 5). However, unlike the best subset selection, the main contribution of forward and backward subset selections to our framework is their ability to rank the variables. For forward stepwise subset selection, the variable that appears earlier is believed to be more important in explaining the OOP costs. For backward stepwise subset selection, it is reversed.

Variable importance ranking: random forests

The random forests approach also provides the rankings of all variables using the two measurements "mean decrease in accuracy" and "mean decrease in node purity" [45]. The measurement "mean decrease in accuracy" is based on how much the accuracy decreases when the variable is excluded. "Mean decrease in node purity" is based on the decrease of Gini impurity when a variable is chosen to split a node. This may associate higher ranking to numerical variables because numerical variables potentially have many split points. The results are provided in Fig. 4. As displayed, *type of insurance coverage*, *race* and *number of physician office visits* are the most leading factors based on "mean decrease in accuracy" measurement. Based on "mean decrease in node purity"

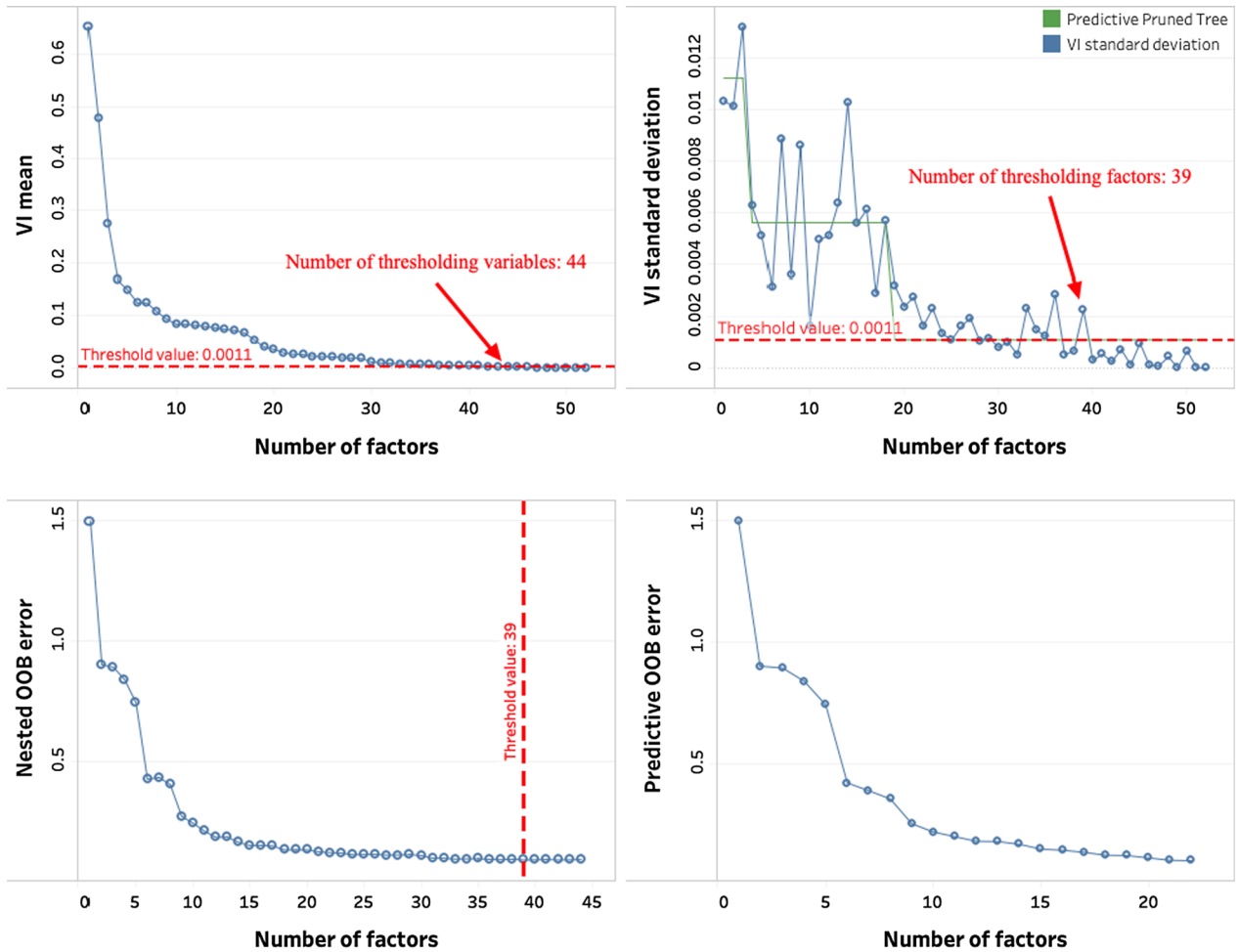

Fig. 3 The variable selection results by the VSURF method. Top graphs illustrate the thresholding step, bottom left and bottom right graphs are associated with interpretation and prediction steps respectively

measurement, *number of physician office visits*, *type of insurance coverage* and *age* rank in the top 3. This observation is consistent with the one output from the best subset selection method. The ranking of *race* in "mean decrease in node purity" measurement is not as high as in "mean decrease in accuracy". That is because *race* has potentially less splits than the other top ranked variables, according to the Gini-based importance.

Votes ranking

In Table 5 below we rank variables that were selected by the best subset selection method based on forward and backward stepwise subset selections as well as random forests method. None of these ranking approaches is perfect, but viewing them altogether allows a comparison of the importance ranking of all variables across all measures. Employing these four rankings, we assign each of the 19 variables listed in Table 3 a score following this rule: If one method ranks a variable in the top five, that variable gets one vote. The score of the variable is then the total count of votes. The more methods that are ranked in the top five, the higher the score received by this variable. The ranking result agrees with the literature [11, 12, 19, 20, 42]. Moreover, our ranking provides useful information on variable importance.

Based on Table 5, *type of insurance coverage* is the most important factor in explaining OOP costs with full score. *Age* receives one less vote, due to the random forests. Hence, OOP costs are affected more by *type of insurance coverage* than by *age*. *Age* is followed by *Asthma*, *family size*, *race* and *number of physician office visits* who equally receive a score of two. Regarding all the ranking methods, forward and backward stepwise subset selections both rank *asthma* as the top one. Compared to other chronic conditions, *asthma* earns the highest score. By using this ensemble learning approach with a voting ensemble, we

Table 4 Variables selected by VSURF and LASSO

Variable	VSURF			LASSO
	Thresholding	Interpretation	Prediction	
Type of insurance coverage	*	*	*	*
Age	*	*	*	*
Asthma	*	*		*
Family size	*	*	*	*
Race	*	*	*	*
Number of physician office visits	*	*	*	*
Family income	*	*	*	*
Primary language not English	*	*	*	*
Sex	*	*	*	*
Joint pain	*	*	*	
Purchased food stamps	*	*	*	*
Occupation	*	*	*	*
Diabetes	*	*	*	*
Work/School/Housework limitation	*	*		*
Marital status	*	*	*	*
Self-employment status	*	*	*	*
Region	*	*	*	*
Serious cognitive difficulty	*	*		*
Cognitive limitation	*	*		*
Employment status				*
Born in the U.S.				*
Year				*
Years in the U.S.				*
Hourly wage level				*
Individual's wage income				*
English proficiency				*

* denotes that the corresponding variable is selected by the method

obtain a list of the most important variables in Table 6 presented in the next section.

Model validation

In this section, we compare the prediction performance of the linear model, random forests, ridge and LASSO according to their test MSEs of the log-odds. The log-odds' test MSEs are estimated by using 5-fold cross validation. More precisely, we derive the smallest test MSEs suggested by the three subset selections, the random forests, ridge and LASSO based on the 26 variables, respectively. We also compare the above test MSEs with the other two models. The first one is the linear regression over the literature recommended factors which is built based on the information gathered in previous studies [4–7, 19, 20, 22, 33, 46, 47]. OOP costs are directly or indirectly affected by insurance status, health status and utilization. This allows us to select six out of the most important variables, as shown in Table 6. The second model is the linear regression over the variables picked by the data-driven solution (see Table 6). These variables are those who receive a score of two or higher based on the rankings in Table 5.

The values of the test MSEs for different approaches are listed below.

From Table 7, we observe that the best subset selection attains its smallest test MSE with 24 variables (44 factors, see Fig. 2). Similarly, the forward and backward stepwise subset selections need almost all the 52 factors (26 variables) to reach their smallest test MSEs. The 25-variable subset suggested by the forward stepwise subset selection has the overall best performance in forecasting, and its test MSE is much less than the 6-variable subset. This indicates that the OOP costs may be determined simultaneously and independently by a large number of factors in a complicated way.

Random forests' performance is better than the two manually input models (literature and data-driven).

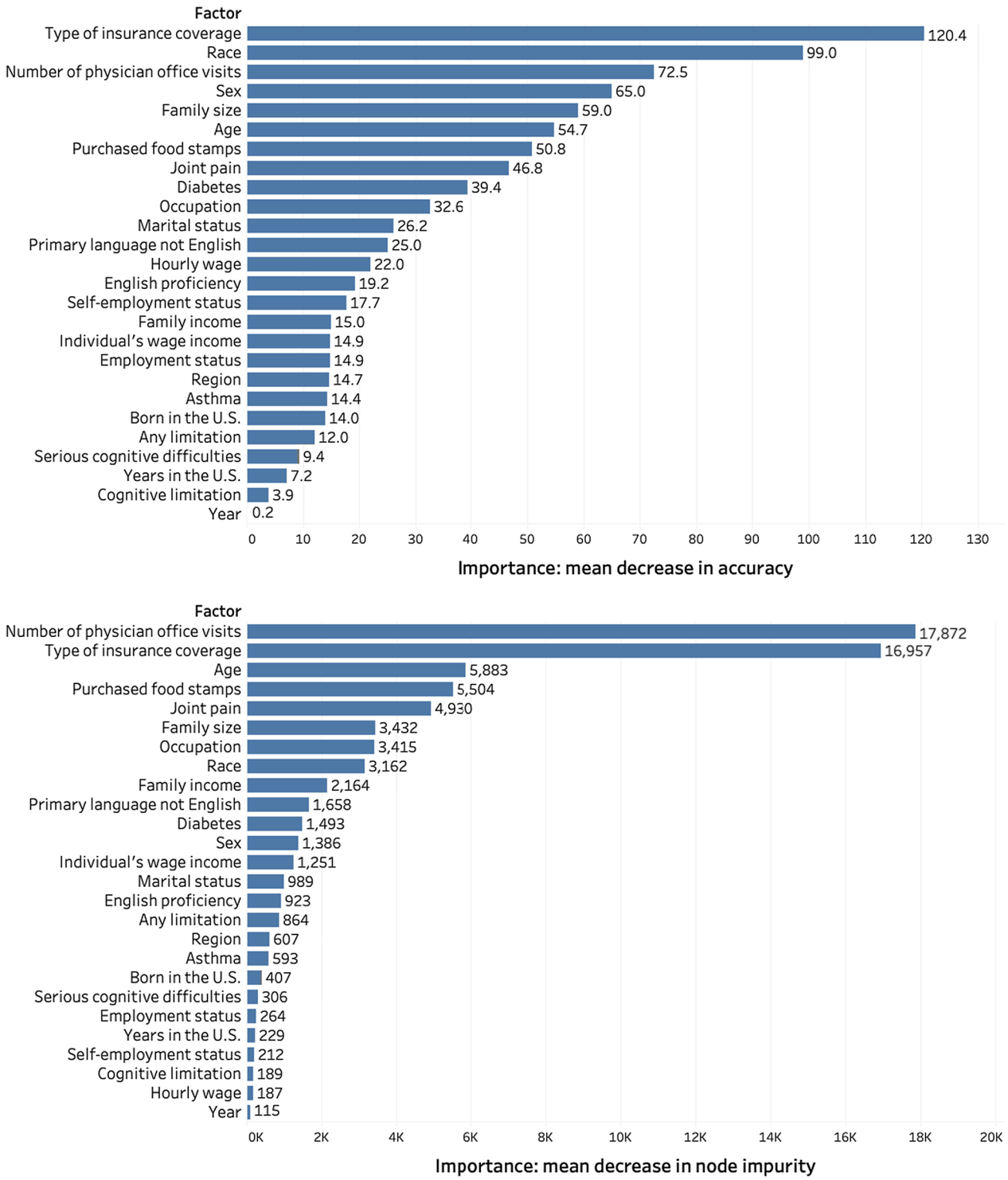

Fig. 4 Rankings of Variables using Random Forests

However, this result is obtained based on 15 variables (see Table 5), which is much larger than 6. Also note that using random forests, it is difficult to explain how each variable impacts the OOP costs.

Table 5 Variable importance rankings

Variable	Ranking				Score
	Forward	Backward	Mean decrease in accuracy	Mean decrease in node purity	
Type of insurance coverage	2 *	2 *	1 *	2 *	4
Age	3 *	3 *	6	3 *	3
Asthma	1 *	1 *	20	18	2
Family size	4 *	7	5 *	6	2
Race	6	5 *	2 *	8	2
Number of physician office visits	20	20	3 *	1 *	2
Family income	5 *	10	16	9	1
Primary language not English	7	4 *	12	10	1
Sex	10	9	4 *	12	1
Joint pain	16	17	8	5 *	1
Food stamp purchased	17	21	7	4 *	1
Occupation	8	6	10	7	0
Diabetes	9	8	9	11	0
Any limitation	11	12	21	16	0
Marital status	12	13	11	14	0
Self-employed status	14	15	15	23	0
Region	15	16	19	17	0
Serious cognitive difficulties	18	14	23	20	0
Cognitive limitation	21	11	25	24	0

* denotes the variable ranks among top five under the corresponding criterion

Table 6 Variables recommended by literature and data-driven solutions

Recommended by literature	Recommended by data-driven solutions
Type of insurance coverage	Type of insurance coverage
Age	Age
Sex	Asthma
Family income	Family size
Race	Race
Number of physician office visits	Number of physician office visits

Table 7 Model validation: test MSEs and the corresponding number of variables

Method	Number of Variables	Test MSE
Best subset selection	24	0.051996
Forward stepwise subset selection	25	0.000113
Backward stepwise subset selection	26	0.000180
Random forests	15	0.098746
Ridge	26	0.004166
LASSO	25	0.054689
Literature recommended	6	0.457194
Data-driven recommended	6	0.371977

According to Table 8, with four different models, the data-driven recommended variables all result in lower training MSE and test MSE compared to the literature recommended ones, which shows the ensemble learning solution performs better in model validation. From the literature recommended variables to the data-driven solution ones, the only difference is the variables *sex* and *family income* are switched to *asthma* and *family size*. We believe the latter pair of variables are more crucial factors of the OOP costs since "having *asthma*" and "increasing *family size*" are straightforward reasons leading to "spending more in health insurance". Moreover, most health plans offer ownership or covered life options such as "single plan" or "joint plan", which have different levels of costs. The variable *family size* directly determines whether the health plan should be "single" or "joint". The result in Table 8 has supported this belief. Finally Table 9 lists the coefficients $\beta_0, \ldots, \beta_6$ of the fitting formula as in (1):

$$\mathbb{P}(y = 1|X = (x_1, \ldots, x_6)) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_6 x_6)}},$$

where $y = 1$ denotes the event "OOP spending over \$1 ,000" and $x_1, \ldots, x_6$ denote the six determinants recommended either by the literature or by the data-driven

Table 8 Comparison of the training MSE and test MSE. MSEs of the four models are calculated using variables (in Table 6) recommended by literature and data-driven solutions

Method	Recommended by Literature		Recommended by Data-driven Solution	
	Training MSE	Test MSE	Training MSE	Test MSE
Linear regression	0.456971	0.457194	0.371514	0.371977
Random forests	0.339891	0.473549	0.311900	0.393755
Ridge	0.458609	0.458814	0.375946	0.376285
LASSO	0.459222	0.459375	0.384613	0.384935

Table 9 Logistic regression coefficients using the recommended variables

Variable	Factor	Literature	Data-Driven
(Intercept)	Numerical	(4.1760)	(6.3214)
Type of insurance coverage	Public health insurance	(1.6850)	(1.7723)
	Uninsured	(0.5132)	(0.5294)
	Private health insurance	–	–
Age	Numerical	0.0305	0.0268
Asthma	No	–	2.9091
	Yes	–	3.1339
	Inapplicable, not ascertained, DK, Refused	–	–
Family size	Numerical	–	(0.2048)
Race	Black	(0.0939)	(0.1132)
	White	0.7511	0.8501
	Multiple races	0.4906	0.5288
	American India/Alaska native	–	–
	Asian/Native Hawaiian/PACFC ISL	–	–
Number of physician office visits	Numerical	0.1306	0.1273
Family income	Numerical	2.8700×10^{-6}	–
Sex	Male	(0.3795)	–
	Female	–	–

– denotes 0; (•) denotes a negative value

solution. Recall that if x_k is numerical, $\beta_k x_k$ is the conventional product of β_k and x_k; if x_k is categorical, $\beta_k x_k$ denotes the additive weight given in (2).

Thresholds of the OOP cost levels

We point out that our variable importance ranking score in Table 5 is based on considering the threshold of the OOP cost levels to be \$1,000 (see Sect. 2). The choice of this threshold is subjective and changing it will influence the ranking result. Therefore, we consider this threshold as an input parameter of our ensemble learning procedure. In Table 10 below we compare the variable importance ranking scores corresponding to the 2 choices of thresholds of OOP cost levels: \$1,000 and \$500. We see that the overall variable rankings differ however the top 3 factors remain the same.

Conclusions, limitations and future research

Considering the rapid growth of OOP costs in the United States, identifying a multitude of OOP costs is crucial for healthcare providers and policymakers to design and implement interventions that reduce disparities in healthcare. This problem involves variable ranking and selection in machine learning. In this paper, we have designed an ensemble learning with voting ensemble for ranking the importance of OOP costs factors in MEPS 2016–2017. The rankings are obtained based on four base

Table 10 Variable importance rankings via different thresholds of OOP cost levels

Variable	Score	
	$1,000	$500
Type of insurance coverage	4	4
Age	3	4
Asthma	2	2
Race	2	1
Family size	2	0
Number of physician office visits	2	0
Primary language not English	1	2
Joint pain	1	1
Family income	1	0
Sex	1	0
Purchased food stamp	1	0

learners: forward, backward stepwise subset selections and two ranking criteria based on random forests. Then we compare the fitting and prediction performance of the six leading factors with that of the six literature supported factors in model validation. Our main contributions are the following:

(1) We obtain that the best subset to explain the behavior of OOP cost levels contains 26 variables out of the 39 variables in MEPS 2016–2017, which indicates that in the real world the OOP costs are impacted by a relatively large number of independent factors in a complicated way.

(2) The top six leading factors selected by our self-designed ensemble learning approach are generally all supported by the literature study. Based on the linear model, our data-driven solution performs slightly better than the six recommended-by-literature variables in terms of prediction.

(3) Our self-designed ensemble learning consists of recently developed tools for variable ranking and selection. With implementation in Python and in *R*, our approach shed some light on applying automatic data-driven tools to deal with data preprocessing, mix-type data correlation detection, variable ranking and selection problems in healthcare data analysis.

Note that there is still room to improve our analysis. For example, splitting the OOP costs values into more than two levels in our analysis will lead to a more sensitive result of variable ranking and selection; developing a way to perform stepwise subset selection for classification problems will make models' quantitative measurements (coefficients, training and test MSEs) more accurate than first transforming the OOP cost levels to log-odds then applying the variable selection approaches. However, this will yield an increase of running time.

Another limitation of this study is the inability to obtain sufficient information. As previously mentioned, MEPS does not provide detailed information on individuals' insurance plans, such as copayments, upper limit and deductibles. For example, whether it is a high-deductible health insurance plan with a risk of high OOP costs, or a low-deductible one with a high monthly premium. Additionally, the information in the MEPS is self-reported; some respondents may refuse to answer, or have no idea on some information in the survey such as chronic disease, employment status. This lack of information may potentially bias our estimates.

Future research is warranted, including examining how determinants influence OOP costs for individuals who have more healthcare needs. In this study, we focused on the working-age adult population, so our results may not be generalizable to older adults with Medicare. Future research with alternative data could also explore how those determinants affect Medicare beneficiaries where the high OOP costs are. It would be also valuable to explore this assumption and investigate in further detail for populations who have a chronic disease (e.g., cancer) that is likely associated with large OOP costs. We leave these exciting topics for further research.

Author details
[1]Claremont Graduate University, Department of Economic Sciences, 150 E. 10th Street, California, Claremont, USA. [2]Claremont Graduate University, Institute of Mathematical Sciences, 150 E. 10th Street, California, Claremont, USA.

Received: 17 October 2020 Accepted: 11 May 2021
Published online: 07 June 2021

References

1. Out-of-pocket spending. Peterson-KFF Health System Tracker. 2017. https://www.healthsystemtracker.org/indicator/access-affordability/out-of-pocket-spending/.
2. Ubel PA, Abernethy AP, Yousuf Zafar S. Full disclosure–out-of-pocket costs as side effects. New Engl J Med. 2013;369(16):1484. https://doi.org/10.1056/NEJMp1306826.
3. Finkelstein EA, Tangka FK, Trogdon JG, Sabatino SA, Richardson LC. Cancer treatment cost in the United States: has the burden shifted over time? Am J Manag C. 2009;15(11):801.
4. Baird K. High out-of-pocket medical spending. among the poor and elderly in nine developed countries. Health Serv Res. 2016;51(4):1467. https://doi.org/10.1111/1475-6773.12444.
5. Gwet P, Machlin SR. Out-of-pocket health care Expenses for non-elderly families by income and family structure, 2015. In: Statistical Brief (Medical Expenditure Panel Survey (US)). 2018
6. Kim NH, Look KA. Effects of the Affordable Care Act's contraceptive coverage requirement on the utilization and out-of-pocket costs of prescribed oral contraceptives. Res Social Adm Pharm. 2018;14(5):479. https://doi.org/10.1016/j.sapharm.2017.06.005.
7. Kang HA, Barner JC. The relationship between out-of-pocket healthcare expenditures and insurance status among individuals with chronic obstructive pulmonary disease. J Pharm Health Serv Res. 2017;8(2):107. https://doi.org/10.1111/jphs.12170.

8. Soni A. Out-of-pocket expenditures for adults with health care expenses for multiple chronic conditions, U.S. Civilian Noninstitutionalized Population, 2014. Statistical Brief (Medical Expenditure Panel Survey (US)). 2017. https://meps.ahrq.gov/data_files/publications/st498/stat498.shtml.
9. Hwang W, Weller W, Ireys H, Anderson G. Out-of-pocket medical spending for care of chronic conditions. Health Affairs. 2001;20(6):267. https://doi.org/10.1377/hlthaff.20.6.267.
10. Paez KA, Zhao L, Hwang W. Rising out-of-pocket spending for chronic conditions: a ten-year trend. Health Affairs. 2009;28(1):15. https://doi.org/10.1377/hlthaff.28.1.15.
11. Hoffman C, Rice D, Sung HY. Persons with chronic conditions. Their prevalence and costs. JAMA. 1996;276(18):1473. https://doi.org/10.1001/jama.1996.03540180029029.
12. Carrier E, Cunningham P. Medical cost burdens among nonelderly adults with asthma. Am J Manag C. 2014;20(11):925.
13. Rodbard HW, Green AJ, Fox KM, S. Grandy for the SHIELD Study Group. Impact of type 2 diabetes mellitus on prescription medication burden and out-of-pocket healthcare expenses. Diabetes Res Clin Pr. 2010;87(3):360. https://doi.org/10.1016/j.diabres.2009.11.021.
14. Opitz D, Maclin R. Popular ensemble methods: an empirical study. J Artif Intell Res. 1999;11:169. https://doi.org/10.1613/jair.614.
15. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning, vol. 1. New York: Springer; 2001.
16. Ho TK. Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. 1995;1:278. https://doi.org/10.1109/ICDAR.1995.598994
17. Cohen JW, Cohen SB, Banthin JS. The medical expenditure panel survey: a national information resource to support healthcare cost research and inform policy and practice. Med Care. 2009;47(7):S44. https://doi.org/10.1097/mlr.0b013e3181a23e3a.
18. Cohen JW, Cohen SB, Banthin JS, The medical expenditure panel survey: a national information resource to support healthcare cost research and inform policy and practice. Medical care. 2009;47(7 Suppl 1):S44. https://doi.org/10.1097/MLR.0b013e3181a23e3a
19. Cylus J, Hartman M, Washington B, Andrews K, Catlin A. Pronounced gender and age differences are evident in personal health care spending per person. Health Affairs. 2010;30(1):153. https://doi.org/10.1377/hlthaff.2010.0216.
20. Cook BL, Manning WG. Measuring racial/ethnic disparities across the distribution of health care expenditures. Health Serv Res. 2009;44(5p1):1603.
21. Derose PK, Bahney BW, Lurie N, Escarce JJ. Review: immigrants and health care access, quality, and cost. Med Care Res Rev. 2009;66(4):355. https://doi.org/10.1177/1077558708330425.
22. Liu S, Chollet D. Price and income elasticity of the demand for health insurance and health care services: a critical review of the literature. Mathematica Policy Research. 2006.
23. Smith-Spangler CM, Bhattacharya J, Goldhaber-Fiebert JD. Diabetes, its treatment, and catastrophic medical spending in 35 developing countries. Diabetes Care. 2012;35(2):319. https://doi.org/10.2337/dc11-1770.
24. Einav L, Finkelstein A. Moral hazard in health insurance: what we know and how we know it. J Euro Econ Assoc. 2018;16(4):957. https://doi.org/10.1093/jeea/jvy017.
25. Ashman JJ, Rui P, Okeyode T. Characteristics of office-based physician visits, 2016. NCHS Data Brief (CDC Stacks). 2019. https://stacks.cdc.gov/view/cdc/62369.
26. Baak M, Koopman R, Snoek HL, Klous S. A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. Comput Stat Data Anal. 2020;152:107043. https://doi.org/10.1016/j.csda.2020.107043.
27. Huh M. The relationships between cognitive function and hearing loss among the elderly. J Phys Ther Sci. 2018;30(1):174. https://doi.org/10.1589/jpts.30.174.
28. Hébert R. Functional decline in old age. CMAJ. 1997;157(8):1037.
29. Long AN, Dagogo-Jack S. Comorbidities of diabetes and hypertension: mechanisms and approach to target organ protection. J Clin Hypertens. 2011;13(4):244. https://doi.org/10.1111/j.1751-7176.2011.00434.x.
30. Visscher TL, Seidell JC. The public health impact of obesity. Annu Rev Publ Health. 2001;22(1):355. https://doi.org/10.1146/annurev.publhealth.22.1.355.
31. Hastie TJ, Tibshirani RJ. Generalized additive models, vol. 43. New York: CRC Press; 1990.
32. Kominski GF, Rasmussen PW, Zhang C, Hassan S, Freund D. Ten years of the Affordable Care Act: Major gains and ongoing disparities. UCLA Center for Health Policy Research. 2020. http://healthpolicy.ucla.edu/publications/search/pages/detail.aspx?PublD=1930.
33. Waidmann T. Estimating the cost of racial and ethnic health disparities. Washington, DC: Urban Institute; 2009.
34. Jacobs EA, Karavolos K, Rathouz PJ, Ferris TG, Powell LH. Limited English proficiency and breast and cervical cancer screening in a multiethnic population. Am J Public Health. 2005;95(8):1410.
35. Harlan LC, Bernstein AB, Kessler LG. Cervical cancer screening: who is not screened and why? Am J Public Health. 1991;81(7):885. https://doi.org/10.2105/AJPH.81.7.885.
36. Hu DJ, Covell RM. Health care usage by Hispanic outpatients as function of primary language. Western J Med. 1986;144(4):490.
37. Cheng EM, Chen A, Cunningham W. Primary language and receipt of recommended health care among Hispanics in the United States. J Gen Intern Med. 2007;22(Suppl 2):283.
38. Planned knee and hip replacement surgeries are on the rise in the us. Blue Cross Blue Shield, Chicago. 2019. https://www.bcbs.com/the-health-of-america/reports/planned-knee-and-hip-replacement-surgeries-are-the-rise-the-us.
39. Barnes PM, Bloom B, Nahin RL. Complementary and alternative medicine use among adults and children; United States, 2007. National Health Statistics Report (CDC). 2008. https://stacks.cdc.gov/view/cdc/5266.
40. Nahin RL, Stussman BJ, Herman PM. Out-of-pocket expenditures on complementary health approaches associated with painful health conditions in a nationally representative adult sample. J Pain. 2015;16(11):1147. https://doi.org/10.1016/j.jpain.2015.07.013.
41. Seuring T, Archangelidi O, Suhrcke M. The economic costs of type 2 diabetes: a global systematic review. Pharmacoeconomics. 2015;33(8):811. https://doi.org/10.1007/s40273-015-0268-9.
42. Karaca-Mandic P, Jena AB, Joyce GF, Goldman DP. Out-of-pocket medication costs and use of medications and health care services among children with asthma. JAMA. 2012;307(12):1284. https://doi.org/10.1001/jama.2012.340.
43. Genuer R, Poggi JM, Tuleau-Malot C. Variable selection using random forests. Pattern Recog Lett. 2010;31(14):2225. https://doi.org/10.1016/j.patrec.2010.03.014.
44. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc: Series B (Methodological). 1996;58(1):267. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
45. Breiman L. Manual on setting up, using, and understanding random forests v3. 1. Statistics Department University of California Berkeley. CA, USA. 2002;1:58.
46. Yu H, Dick AW, Szilagyi PG. Does public insurance provide better financial protection against rising health care costs for families of children with special health care needs? Med Care. 2008;46(10):1064. https://doi.org/10.1097/MLR.0b013e318185cdf2.
47. Shen C. Determinants of health care decisions: insurance, utilization, and expenditures. Rev Econ Stat. 2013;95(1):142. https://doi.org/10.1162/REST_a_00232.