Economic Modelling 130 (2024) 106564

Contents lists available at ScienceDirect

Economic Modelling

journal homepage: www.journals.elsevier.com/economic-modelling

Addressing unanticipated interactions in risk equalization: A machine learning approach to modeling medical expenditure risk

I. Ismail [a,*], P.J.A. Stam [a,b], F.R.M. Portrait [a], A. van Witteloostuijn [a,c], X. Koolman [a]

[a] School of Business and Economics, Ethics, Governance and Society, Vrije Universiteit Amsterdam, De Boelelaan 1085, 1081, HV Amsterdam, the Netherlands
[b] Equalis Strategy & Modeling B.V, Jaarbeursplein 6, 3521 AL, Utrecht, the Netherlands
[c] Antwerp Management School, Boogkeers 5, BE-2000, Antwerpen, Belgium

ARTICLE INFO

Handling editor: Sushanta Mallick

JEL classification:
C55
I11
I13
G22

Keywords:
Risk equalization
Machine learning
Risk selection
Model fitting
Health insurance

ABSTRACT

Adverse selection harms market efficiency and access to essential services, particularly for disadvantaged groups. Risk equalization policies attempt to mitigate this by compensating agents for risk disparities, but often fall short of addressing interactions between risk factors. Using health insurance data from the Netherlands, we present a machine learning approach to capture unanticipated interactions that impact medical expenditure risk. We compare our novel approach to a state-of-the-art statistical model. We find that our approach explains an additional 1.5% of medical expenditure, equivalent to 571 million euros over all individuals in the Dutch market. In particular, this translates into better compensation for low- and high-cost groups that are especially vulnerable to adverse selection. These findings confirm the significance of risk factor interactions in explaining medical expenditure risk, and support the adoption of machine learning alongside statistical models to further mitigate selection incentives in risk equalization policies.

1. Introduction

Risk equalization (RE) is an important regulatory instrument in markets that face adverse selection problems, a prominent example being health insurance (Aguirre and Beitia, 2017). Adverse selection encompasses actions taken by both insurers and consumers to exploit unpriced risk variation, leading to adverse market outcomes. In health insurance, selection incentives are known to undermine access, service quality, and overall market efficiency (Layton et al., 2017; Van de Ven et al., 2017, 2023). Similar challenges also arise in other economic settings facing adverse selection, such as local government service provision and education. In these settings, RE mechanisms like fiscal equalization and weighted student funding are used (Bos, 2013; Eichhorst, 2007; Ladd and Fiske, 2011; Roza et al., 2021). These policies aim to minimize adverse selection problems and level the playing field by compensating agents for the risk disparities they encounter. Improving the effectiveness of RE in these public-oriented markets would benefit society.

Our key argument is that machine learning can help regulators improve RE when used complementary to prevailing statistical methods. One recognized issue with statistical RE models is their inability to address the impact of unanticipated – yet relevant – interaction effects between risk factors. This leads to inadequate compensations that still leave room for selection incentives (Layton et al., 2017; Van de Ven et al., 2017, 2023). Machine learning's ability to algorithmically learn from data without predefining a data generation process solves this limitation and can help address risk factor interactions more effectively (Breiman, 2001b). Taking the example of health insurance, we introduce a population-wide machine learning approach to RE modeling and contrast our novel approach to standard statistical modeling. We discuss where these methods may complement each other in light of policy requirements like transparency and provide recommendations for future RE implementations.

While relevant in many sectors, the use of RE is most sophisticated in health insurance markets that employ a form of regulated competition to manage efficiency and equity (Van de Ven and Ellis, 2000). This includes basic health insurance in Germany, Israel, and the Netherlands, supplementary health insurance in Australia and Ireland, and Medicare Advantage and the Affordable Care Act's Marketplaces in the United States. In all these examples, regulators use statistical models to

[*] Corresponding author. De Boelelaan 1085, 1081, HV Amsterdam, the Netherlands.
E-mail address: i.ismail@vu.nl (I. Ismail).

https://doi.org/10.1016/j.econmod.2023.106564
Received 2 September 2022; Received in revised form 10 October 2023; Accepted 13 October 2023
Available online 19 October 2023

compensate insurers for the medical expenditure risk of their planholders. In the past three decades, RE models have been gradually improved by adding new risk classes or refining the definition of existing ones. However, despite these efforts, even state-of-the-art models, like those used in the United States and the Netherlands, remain prone to under- and overcompensation (McGuire et al., 2020; Van de Ven et al., 2017).

Typically, RE models apply linear regression to estimate medical expenditure based on demographic, morbidity-related, and cost-based risk factors. Interactions effects are only moderately included in such models. Some two-way interactions, like between age and sex, have been well documented and have since been included in RE implementations. Higher-order interactions, specifically among diagnostic, pharmaceutical or cost-based risk classes, are however often left unaddressed (Van Veen et al., 2017). While theoretical literature on higher-order interactions is limited, such effects may help further explain the heterogenous medical expenditure observed in smaller subgroups with complex risk profiles, as in the case of co-morbidity (Oskam et al., 2023).

Machine learning involves a modeling tradition very different from classic statistics, as explained by Breiman at the beginning of this century (2001b). In contrast to classic statistics, machine learning gives full leeway to the data, with the 'machine' learning the actual data generation process iteratively. The essential difference lies in the (non-)use of a priori assumptions regarding the form and nature of the full data generation process. As a result, machine learning can flexibly accommodate any higher-order interaction that improves capturing the data generation process. Recent studies have explored machine learning applications to improve RE, but mostly focused on variable selection or evaluated machine learning models on limited samples that were not representative at the population level (Iommi et al., 2022; McGuire et al., 2021; Van Veen et al., 2017). Critically, these samples may omit the smaller subgroups with complex risk profiles for whom interactions may be most relevant.

We address this gap in literature by presenting a population-wide machine learning approach to capture unanticipated risk factor interactions and determine their impact on medical expenditure risk. Our contribution consists of three aspects. First, by using population-wide data, we prevent sampling bias and can explore the ability of machine learning methods to detect interaction effects across all possible subgroups. Second, we benchmark our approach against the Dutch RE model, which is considered state-of-the-art internationally (McGuire et al., 2021; Van de Ven et al., 2023). This allows us to quantify the added value of machine learning and, specifically in the context of this study, its ability to capture interactions that were not anticipated in the Dutch RE model. Finally, we impose constraints in line with substantive criteria that regulators typically encounter to replicate real-world conditions, and place particular emphasis on transparency requirements.

Our findings allow insight in the effectiveness of machine learning in addressing unanticipated risk factor interactions, their impact on individual medical expenditure risk, and how this translates into the compensation of low- and high-risk subgroups in the population. We set a threshold for the extent to which regulators may improve their RE system by adopting machine learning, and provide guidance on the process. These findings should be relevant to the numerous health systems that rely on RE to preserve societal goals. More generally, we add to this journal's conversation regarding nationwide health system issues across many countries, such as medical expenditure risk (Du, 2023, on the US), health insurance reform (Cui et al., 2021, on China), access inequity (Brown et al., 2014, on Turkey), drivers of expenditure (Murthy and Okunade, 2016, on the US), and price dispersion (Oliva and Carles, 2013, on Catalonia), introducing the issue of RE as a regulatory instrument.

The structure of the article is as follows. Section 2 provides further technical background. Section 3 describes the data, model development, model evaluation, and functional analysis performed. Section 4 presents our findings in terms of model performance and gives an overview of important model parameters and their effects. Section 5 includes a discussion of key findings, limitations, and the implications for regulators and insurers.

2. Technical background

This study leverages two established machine learning algorithms: random forests (RFs) and gradient boosting machines (GBMs). As shown by the emerging stream of research in this journal, these tree-based algorithms can be used to model a wide arrange of issues, including earthquake losses (Gu et al., 2023), IPO failure risk (Colak et al., 2022), and micro-simulation (Hughes et al., 2022). From all machine learning methods, tree-based models have also been most extensively applied in RE (Iommi et al., 2022; Van Veen et al., 2017). Regression trees, which form the base of tree-based models, offer a non-parametric modeling technique that can identify higher-order interactions between predictors. While relatively interpretable, singular regression trees can be prone to overfitting (Hastie et al., 2009). RFs and GBMs are two ensemble approaches that address this issue by combining multiple regression trees in a parallel or sequential structure, while retaining the ability to detect higher-order interactions (Friedman, 2001). An important advantage of these methods, particularly in policy context, is their comprehensive research process pipeline, which includes model interpretability metrics comparable to those we are familiar with in classic statistics.

Our approach is centered on leveraging the ability of RFs and GBMs to automatically capture higher-order interactions between risk classes, while regularizing model complexity to ensure robustness of findings. We expect that allowing for more and higher-order interactions, provided that they are structural, will improve the prediction of medical expenditure in subgroups with complex risk profiles, which will ultimately result in overall better compensation of insurers. To test whether our machine learning approach results in better RE, we apply a thorough out-of-sample evaluation at the individual and subgroup level based on traditional RE performance criteria (Geruso and Mcguire, 2016). However, the ensemble structure of the applied methods means that model transparency requires further inspection. To understand the conceptual differences between the machine learning models and the existing linear RE model, we turn to modern interpretable machine learning (IML) techniques, such as variable importance and marginal effect analyses (Friedman, 2001). Note however that the primary focus in RE is accurate prediction; and that for this reason, we put more emphasis on model performance.

Finally, to ensure comparable conditions in our evaluation, we take substantive criteria often applied to RE models into consideration. These include fairness, appropriateness of incentives, and feasibility (Van de Ven and Ellis, 2000). Fairness refers to the solidarity between low-risk and high-risk enrollees, appropriateness of incentives involves distortions or undesirable behavioral responses to incentives created by the RE model, and feasibility concerns transparency of the model plus the practicality of development and implementation. Moreover, a distinction is made between factors for which solidarity is desired, known as S (ubsidy)-type factors, and N(on-subsidy)-type factors for which this is not desired (Schokkaert et al., 1998; Van de Ven and Ellis, 2000). S-type factors include aspects such as health, age, and sex, while N-type factors concern risk adjusters considered as the responsibility of insurers such as differences in supply, price, or practice patterns. Regulators impose restrictions on the risk adjusters used in RE models to reflect these criteria. We adhere to this by taking all data as given, and not introducing any alterations to the input data.

3. Methods

3.1. Dataset and variables

This study makes use of the so-called 'Overall Test 2018' administrative dataset that was constructed for the development of the Dutch RE model of 2018 (year *t+3*). The dataset contains individual-level information on pre-defined risk classes measured in 2014 (year *t-1*) and somatic medical expenses incurred in 2015 (year *t*) of all Dutch citizens (N = 17,004,068). The data was primarily sourced from insurance claims data, tax authorities, and the social benefits registration service. Note that it is standard practice to estimate the prospective Dutch RE model (year *t+3*) based on historical data (year *t-1* and *t*) since more recent population-level data is usually not available. All variables are taken as given to warrant consistency with the Dutch RE model, since it forms a point of comparison throughout this study. The response variable concerns the somatic medical expenses measured per individual in euros. The risk adjusters include pre-defined risk classes specified as binary variables indicating whether an individual belonged to that risk class. A description of the Dutch RE model's risk classes is provided in Table 1.

3.2. Data preparation

The dataset ($N = N_1 + N_2$) was randomly split using a 70:30 ratio into a training set R ($N_1 = 11,901,755$) and a test set S ($N_2 = 5,102,313$). The training set was used to develop the models and the test set to evaluate the models. As is standard in machine learning's research, we opted for an out-of-sample assessment to penalize models for potential overfitting (Hastie et al., 2009). The training set was vertically aggregated to reduce the computational burden of training the machine learning models. This means that individuals with the same set of risk classes were combined into one row to yield a dataset in which each observation represented a unique combination of risk classes (N_1 = 1,355,008). Furthermore, the mean somatic medical expenses was computed for each row and a weight was assigned that represents the number of individuals within a row. The test set was used to estimate the out-of-sample predictive performance on data unseen during training.

3.3. Methods of analysis

First, we fitted a no-intercept OLS model to simulate the Dutch RE model of 2018 using the somatic medical expenses as the response Y and the risk classes X_j as predictors on training set R. The linear relationship between Y and the predictor set $X = \{X_1, \ldots, X_J\}$ assumes a model of the form $Y = f(X) + \varepsilon$ with ε a mean-zero random error term and a function f, given by

$$f(X) = \sum_{j=1}^{J} \beta_j X_j, \tag{1}$$

where Y is a $(N_1 \times 1)$ vector $(y_1, \ldots, y_{N_1})^T$, X_j a $(N_1 \times 1)$ vector, β_j the parameters to be estimated, and J the total number of risk classes as described in Table 1. It is common practice in the Netherlands to include all age and gender classes in Equation (1) and to apply an identifying constraint during the estimation phase – i.e., the expenses predicted by these classes should sum up to the total expenses.

Second, we developed two alternative machine learning RE models on training data R using the RF and GBM algorithms without making any alteration to the input variables. We made use of the *Ranger* implementation of the RF algorithm and the *XGBoost* implementation of the GBM algorithm, which effectiveness has been shown in economic analysis before (Alanis, 2022; Carmona et al., 2019).

Table 1
Risk classes as defined in the 'Overall Test 2018' administrative dataset.

Risk classes	Description
Age and gender (A&G: 42 classes)	21 age classes for males and females with a 5-year interval except for the following deviant age groups: 0 years (two classes), 15–17 years, 18–24 years and >90 years.
Pharmaceutical cost groups (PCG: 34 classes)[a]	34 classes of PCGs to which an individual can be assigned based on the extramural medicines that were used. A threshold of 180 daily dosages is used, otherwise an individual is assigned to the 'no PCG' reference class. Individuals can be assigned to multiple PCGs.
Primary diagnostic cost groups (pDCG: 16 classes)	16 classes of primary diagnoses to which individuals can be assigned based on hospital admissions with a reference class for individuals without any such diagnosis. Assignment to multiple pDCGs is not possible.
Secondary diagnostic cost groups (sDCG: 8 classes)	8 secondary diagnostic cost classes that follow the same logic as the pDCGs. sDCGs reflect secondary diagnoses and intend to improve compensation in case of multimorbidity.
Durable medical equipment groups (DME: 11)	11 classes based on the usage of durable medical equipment, including a class for those who have not used any.
Source of income and age (SI: 25 classes)[b]	6 source of income classes (social security payments, full disability payments, miscellaneous disability payments, students, self-employment and higher-educated) interacted with 4 age groups (18–34, 35–44, 45–54, 55–65) and two additional classes for <18-year-olds and >65-year-olds.
Region (10 classes)	10 clusters of ZIP code areas that are based on commonly shared risk characteristics. These include healthcare supply, socioeconomic circumstances, and residual health differences. The clustering is not necessarily related to geographic proximity.
Socioeconomic status and age (SES: 12 classes)	4 socioeconomic classes (very low, low, mid, and high income) interacted with three age groups (0–17, 18–64, 65+).
Persons per address and age (PPA: 12 classes)	4 PPA classes (permanently institutionalized, temporarily institutionalized, living alone and miscellaneous) interacted with three age groups (0–17, 18–64, 65+).
Multiple-year high-cost groups (MHC: 9 classes)	6 classes for those with consecutive high costs (top 15%, 10%, 7%, 4%, 1.5% and 0.5% costs) in the previous three years. One class for those who were two years within the top 10% and once *not* in the top 15%. Two classes for those without high costs in consecutive years (bottom 85% or bottom 70%).
Physiotherapy cost groups (PTCG: 5 classes)	4 classes based on physiotherapy claims and one reference class for those without any claims.
Costs nursing and care year t-1 (CNC: 8 classes)	7 classes based on the costs of nursing and caring in the previous year. The highest risk class is split into two age groups (<18 or ≥ 18).

[a] The only risk adjuster for which the risk classes are not mutually exclusive.
[b] Number of classes does not add up as interaction with age does not hold for every source of income group.

3.3.1. Regression analysis of aggregate data

At the aggregate level, each row in the data contains a unique combination of risk classes, a weight representing the number of individuals with that set of risk classes, and the computed mean somatic medical expenses. For standard OLS applied to a linear regression model with continuous y and categorical variables, such as the binary risk classes in our study, parameter estimates calculated using aggregate data should align with those calculated using individual data (Nicoletti and Best, 2012). No information on the distribution of expenses other than the mean is necessary to fit the RE models at the aggregate level. Since the parameter estimates of both individual and aggregate regression models are the same, it follows that the minimized sum of squared errors (SSE) at the aggregate and individual level is identical, where SSE based on individual observations is defined as

$$SSE = \sum_{i \in R} (y_i - \hat{y}_i)^2, \tag{2}$$

with y_i being the observed value for an individual i and $\hat{y}_i$ the predicted value for that individual i. The mean squared error (MSE) can be derived by dividing the SSE by the number of rows. Note that information on variances would be needed to calculate confidence intervals for these parameter estimates, which is lost due to vertical aggregation. However, we do not test for statistical significance or calculate confidence intervals as this would be uninformative considering the size of our data, and unnecessary for the predictive evaluation we aim for in the current study.

3.3.2. Tree-based models explained

Regression trees are non-parametric machine learning algorithms that use a set of if-then rules to partition data into subregions that are more homogenous regarding a response variable (Breiman, 2001a). This is achieved by searching the optimal predictor and corresponding value that partitions the data such that the within-error of the resulting subregions is minimized. The splitting process begins with the entire training set, and searches the predictor and corresponding split value that partitions the training data into two groups (R_1 and R_2) such that the overall sums of squared errors (SSE) are minimized, given by

$$SSE = \sum_{i \in R_1} (y_i - \hat{y}_{R_1})^2 + \sum_{i \in R_2} (y_i - \hat{y}_{R_2})^2, \tag{3}$$

where $\hat{y}_{R_1}$ and $\hat{y}_{R_2}$ are the averages of the training set outcomes within groups R_1 and R_2, respectively (Hastie et al., 2009). Subsequently, within each of the groups R_1 and R_2, this method searches the next predictor and split value that as a pair best reduces the SSE. The splitting process is reiterated within each resulting subregion until a tree with a minimum error rate is constructed, i.e., groups $R_1, \ldots, R_M$ are formed that minimize the overall SSE given by

$$SSE = \sum_{m=1}^{M} \sum_{i \in R_m} (y_i - \hat{y}_{R_m})^2, \tag{4}$$

where $\hat{y}_{R_m}$ is the average outcome for the training observations within group R_m and M is the total number of groups (Hastie et al., 2009). The resulting regression tree, $f(X)$, can be summarized as

$$f(X) = \sum_{m=1}^{M} c_m * I(X \in R_m), \tag{5}$$

where R_m indicates subregions of the predictor space, I(.) a function that returns 1 if its argument is true and 0 otherwise, and c_m the estimated value of the response variable in subregion R_m (Hastie et al., 2009). As in the case of standard OLS (see Subsection 3.3.1), no information on the distribution of the outcome variable other than the mean is needed to fit single regression trees or ensemble variants. Therefore, we can take advantage of vertical aggregation in this case, too.

Regression trees naturally identify interaction effects of any order, as each asymmetrical branch in the regression tree is perceived as a local interaction between the involved predictors. A significant advantage is thus that interaction effects do not have to be specified manually. This illustrates machine learning's flexibility since the nature of the data generation process is learned ex post rather than assumed ex ante. After all, due to the logic of their structure, regression trees do not rely on a pre-specified functional form in any way. Single tree-based models can however be unstable, since they tend to overfit as tree complexity increases. As a result, the predictive performance of such models is generally suboptimal. The RF and GBM algorithms solve the lack of generalizability by applying different ensemble approaches, which we explain below.

3.3.3. Random forests

RFs are an ensemble approach to regression trees that incorporates bootstrap aggregation to minimize the variance of single tree models (Breiman, 2001a). In bootstrap aggregation, random cases are sampled with replacement from a given dataset to yield B bootstrap samples. A full-grown regression tree is trained on each bootstrap sample and the RF prediction is obtained by averaging the estimations results across all independent trees.

Each tree is constructed as described above, but only a random selection of predictors is considered in the search to find each split. The induced randomness reduces the correlation between the independent trees, and thereby further improves the predictive capacity of the RF. The independent trees retain their tendency to overfit, but the bias of each tree is eliminated when averaging out the predictions. An RF model can be summarized as

$$f(X) = \frac{1}{B} \sum_{b=1}^{B} f_b(X), \tag{6}$$

where $f(X)$ represents the averaged prediction of the RF, B the number of bootstrap samples, and $f_b(X)$ the prediction of the regression tree fitted on bootstrap sample b.

3.3.4. Gradient boosting machines

The GBM algorithm forms an ensemble approach that aims to reduce the bias of regression trees by applying the general idea of gradient boosting (Friedman, 2001). Gradient boosting combines multiple 'weak learners' sequentially to gradually reduce the error of an ensemble model. Regression trees form an ideal weak learner since their complexity, i.e., interaction depth, can be regularized. Restricting the interaction depth also reduces overfitting.

Gradient boosting is initialized by fitting a weak learner, $f_0(X)$, and predicting the response. Note that this initial learner often concerns a tree model with no partitions – i.e., the mean of the response. It is assumed that the prediction of $f_0(X)$ can be improved by explaining the residuals of the learner. A second learner is therefore fit to predict these residuals, given by

$$r_1(X) = y - f_0(X), \tag{7}$$

where $r_1(X)$ is a tree model approximating the error between the observed response y and the predictions of $f_0(X)$. The two learners are combined in an additive manner to form an ensemble, which can be summarized as

$$f_1(X) = f_0(X) + r_1(X). \tag{8}$$

The ensemble is extended by adding several learners until the error between the predicted and observed values is minimized. The GBM then evolves as follows:

$$f(X) = f_0(X) \rightarrow f_1(X) = f_0(X) + r_1(X) \ldots \rightarrow f_p(X) = f_{p-1}(X) + r_p(X), \tag{9}$$

where p is the number of learners used in the GBM. The predictions of the initial learner are iteratively updated using the predictions of subsequent learners, and the final estimates of the ensemble are obtained by summing the prediction of all learners.

3.4. Model development

The RF and GBM algorithms, while relatively robust, both require hyperparameter optimization. This is another critical step in machine learning's research process pipeline, where the aim is to optimize the model's predictive performance on unseen data. The hyperparameters dictate the structure of the model and the rules through which the algorithm learns to model the response, as described in Table 2. The optimization process we applied consisted of three steps that will be elaborated on in the following subsections. First, we defined a set of

Table 2
Hyperparameters of the RF and GBM algorithms.

RF	GBM
N. trees: number of independent trees grown on bootstrap samples of the original data.	N. trees: number of trees that are used sequentially to minimize the residuals.
k (mTry): number of random predictors considered at each split within trees.	Tree complexity: maximum number of splits allowed within a tree, i.e., the interaction depth. Learning rate (shrinkage): a weight dictating the extent to which each tree is allowed to influence to final model. Minimum child weight: minimum number of training set samples in a node to commence splitting. Gamma: minimum improvement in error reduction to commence splitting. Column fraction: the fraction of random predictors to be considered within each tree. Subsample fraction: the fraction of random observations to be sampled for each tree

hyperparameter values to consider through a grid or random search. Second, a 'candidate' model was fit for each combination of hyperparameter values. Third, the performance of each candidate model was assessed through adaptive k-fold cross-validation. The best performing candidate model was used for the out-of-sample evaluation.

3.4.1. *Hyperparameter optimization*

Hyperparameter values can be optimized through several search strategies. In this study, we applied a combination of grid and random searches (Hastie et al., 2009; Ozden and Guleryuz, 2022). Our optimization strategy was informed by the number of hyperparameters and the sensitivity of performance to hyperparameter selection of each algorithm, but also took practicality into account considering the sample size used for model development.

In a grid search, the hyperparameter values to be considered are specified manually and stored in a grid. This grid is then used to specify candidate models for each possible combination of values. In a random search, multiple sets of hyperparameter values are randomly drawn to form the candidate models. Random searches are considered more efficient than grid searches, but the latter are a viable option when the number of hyperparameters is low or if a priori insight on the optimal values is available.

In the case of the RF hyperparameters, it is known that increasing the value of N.trees generally yields better results, while the effect of the *k* parameter is still debated (Hastie et al., 2009). We identified the optimal value of *k* using a grid search, and then gradually stepped up the number of trees. The grid contained values ranging from a 0.1 fraction of the total number of predictors to a 0.5 fraction, with a sequential increase of 0.1 in between. The N. trees parameter was initially set at 500 trees and increased by steps of 100 until the model's performance plateaued.

For the GBM, the number of hyperparameters and the related computational burden is relatively high. Amongst these hyperparameters, some empirical studies show that the number of trees, learning rate, and tree complexity are relatively influential, while randomization parameters such as column- and subsample fraction have a smaller effect on performance (Hastie et al., 2009). We therefore applied a two-step strategy. First, we performed a random search with 30 trials to identify a smaller search domain of well-performing hyperparameter values. Then we used a grid search to further optimize the number of trees, learning rate, and tree complexity hyperparameters. The grid contained values in proximity of the hyperparameter values of the best-performing candidate model identified in the random search.

3.4.2. *Cross-validation*

To select the optimal final hyperparameter tuple, each candidate model constructed through the grid or random search was assessed using adaptive k-fold cross-validation (Zhang and Yang, 2015). In k-fold cross-validation, the original training data is randomly divided into k subsets. Each of the k subsets acts as a simulation test set, and the other k-1 subsets figure as a simulation training set. A candidate model is fit on each of the generated training sets and then tested on the corresponding test sets. The error of all test model iterations is averaged to give an interim measure of performance – the R-squared in our case – of the candidate model. A futility analysis is performed to evaluate candidate models at each resampling step; inferior models are eliminated prematurely to streamline the cross-validation process. To deal with the computational burden of population scale modeling, we restricted the number of folds to k = 5. The interim measure of predictive performance was used to select the final RF and GBM models.

3.4.3. *Regularization of overfitting*

The approach to regularizing overfitting consisted of three layers. First, tuning the hyperparameters allows us to balance the complexity of the models and restrain overfitting. Second, cross-validating all candidate models, as described in the optimization process above, offers the opportunity to select the models that perform the best on 'new' data unseen by the trained algorithm. Third, penalizing all models for possible overfitting by performing an out-of-sample evaluation on the test data implies that we can get a better measure of how the final models would perform on external data.

3.4.4. *Transformation of predicted expenses*

Turning to our application to RE in Dutch health insurance regulation, reservation of the mean of the actual expenses is essential to RE models as the sum of money to be distributed across insurers should be equal to the incurred expenses. In contrast to classic OLS regression, the RF and GBM algorithms are non-parametric methods; therefore, they are not necessarily mean-preserving (Breiman, 2001a; Friedman, 2001). Hence, the predictions of the RF and GBM were rescaled to match the mean of the actual expenses. To do so, we derived a so-called smear factor for both algorithms by dividing the actual expenses by the fitted values in the training dataset. These smear factors, 1.01 and 0.92 for the RF and GBM, respectively, were used to rescale the predictions in the test dataset prior to the evaluation of the model fit such that the mean of the predictions matched the mean of the actual expenses.

3.5. *Model evaluation*

The RF, GBM and OLS models were evaluated on overall predictive accuracy and the net compensation of subgroups on test set *S* using traditional RE criteria (Geruso and Mcguire, 2016). At the individual level, this includes the R-squared (R^2), Cummings prediction measure (CPM), and the mean absolute prediction error (MAPE). At the subgroup level, this involves the MAPE and the mean equalization result (MER).

At individual level, the R^2 gives a standardized measure that ranges from 0 to 1 that indicates the proportion of the variance in the response variable that can be explained by a predictive model. The R^2 was calculated as the square of the correlation coefficient between the observed and predicted values, as suggested by Hocking (2003). The R^2 gives larger weight to large errors and is therefore relatively dependent on individuals with large expenses. The R^2 is derived as

$$R^2 = \left(\sum_{i \in S} (\frac{\widehat{y}_i - \overline{\widehat{y}})(y_i - \overline{y})}{\sqrt{\sum_{i \in S} (\widehat{y}_i - \overline{\widehat{y}})^2 \sum_{i \in S} (y_i - \overline{y})^2}} \right)^2, \tag{10}$$

where $\overline{y}$ is the mean of the observed response variable, y_i the observed value for an individual, $\overline{\widehat{y}}$ the mean of the predicted values, and $\widehat{y}_i$ the predicted value for an individual.

The CPM indicates the proportion of the sum of absolute deviations from the mean that can be explained by a predictive model. The CPM is similar to the R^2 in that it is a standardized measure that ranges from 0 to 1. Unlike the R^2, the CPM assigns equal weights to small and large errors. The CPM is described by Equation (11), where all parameters retain the meaning stated for Equation (10):

$$CPM = 1 - \frac{\sum_{i \in S} |y_i - \hat{y}_i|}{\sum_{i \in S} |y_i - \bar{y}|} . \tag{11}$$

The MAPE gives an absolute measure of the average deviation of the predicted values from the observed values. The MAPE is similar to the CPM in that it is expressed in absolute terms but differs from the CPM and R^2 in that it is not a standardized measure. The MAPE at the individual level, where all parameters retain the meaning stated for Equation (10), is derived as

$$MAPE = \frac{1}{N_2} \sum_{i \in S} |y_i - \hat{y}_i| . \tag{12}$$

At the subgroup level, the MAPE gives an absolute measure of the average deviation of the predicted values from the observed values summed over all subgroups that are included in the RE model. The MER gives the mean difference between the predicted expenses and the actual expenses – i.e., the net compensation for selected subgroups that are *not* already included in the RE model. The MER describes 'unpriced risk heterogeneity' and therefore provides information on incentives for insurers to engage in risk selection of particular subgroups (Withagen-Koster et al., 2018). The MER is reported for two separate sets of subgroups based on actual and historical expenses, respectively.

The first set of subgroups for which the MER is calculated, concerns population deciles ordered by expenses incurred in year *t*. Since these population deciles are based on actual expenses, deviations include incidental expenses as well. Incidental expenses are considered as typical insurance risk. However, the magnitude of undercompensation by even sophisticated RE models suggests that part of such deviations may also reflect omitted variables bias (Ellis and McGuire, 2007). The addition of higher-order interactions among risk classes that follows from using a machine learning approach may help address this. Therefore, the gap between predicted and actual expenses per decile is a relevant model performance indicator.

The second set of subgroups concerns the top 15%, the middle 70%, and the lowest 15% of the population ranked by expenses in year *t-3*. These subgroups are defined using expenses from three years prior to the data used in the model construction to attenuate the impact of incidental expenses. Therefore, the MER in these subgroups represents structural under- and overcompensations better. The definition of these subgroups is in line with conventions of the Dutch regulator. Note however that studies outside of the Netherlands may instead report subgroups based on prior year *t-1* expenses for the same purpose of filtering out incidental expenditure.

3.6. Model interpretation

To help interpret our models considering transparency criteria, which are critical for regulatory agencies, we add functional analyses on variable importance and marginal effects of risk classes. In classic OLS regression, the marginal effect of each predictor is expressed through a regression coefficient, and the importance of each predictor in predicting the response variable can be assessed through standardized coefficients (Siegel, 2016). To obtain similar insights in the RF and GBM, we make use of IML methods introduced by Friedman (2001). Variable importance is here expressed as the number of times a predictor is used to perform a split within the regression tree, weighted by the improvement in error reduction resulting from each split averaged over all trees within the ensemble. The resulting measure provides insight in the relative importance of risk classes, similar to standardized coefficients for classic OLS regression models. Marginal effects are computed by estimating the response in the somatic medical expenses to changes in a sole predictor while holding the effect of the remaining predictors constant. Since all risk classes are binary coded, the computed marginal effects can be interpreted similarly to the regression coefficients of the classic OLS regression model. Whereas the marginal effects analysis solely reflects the main effect, variable importance also accounts for interactions with other risk classes, if they exist. This gives complementary insight into which risk classes are given different weight in the three modeling approaches, and whether these differences relate to varying main effects or potential interactions.

4. Results

4.1. Population descriptives

The descriptive statistics of the total population, training sample, and test sample are given in Table 3. The mean somatic medical expenses of the total population were €2239. The percentage of males was 49.5, and the median age group was 45–49. 19.6% was classified in a PCG and 4.5% belonged to multiple PCGs. 10% was classified in a pDCG and 4.5% in a sDCG. 3.5% was classified in a CNC class, 2.1% in a DME, 5.9% in an MHC, and 26.2% in at least a PCG, pDCG, sDCG, CNC, DME or MHC. The mean somatic medical expenses, age and gender distribution, and prevalence of classes in the training and test sample were nearly identical to that of the total population.

4.2. Prediction of actual expenses

The predictive performance metrics of the OLS, RF and GBM models are reported in Table 4. At the individual level, the RF outperforms the OLS model with a 1.5 percentage point improvement on the R^2, 0.8

Table 3
Descriptive statistics of the population.

	Total population ($N =$ 17,004,068)	Training sample ($N_1 =$ 11,901,755)[a]	Test sample ($N_2 =$ 5,102,313)
Mean somatic medical expenses in euros (SD)	2239 (7823)	2239 (7831)	2242 (7804)
Male (%)	49.5	49.5	49.5
Median age group	45–49	45–49	45–49
Classified in a PCG (%)	19.6	19.7	19.6
Classified in multiple PCGs (%)	4.5	4.5	4.5
Classified in a pDCG (%)	10.0	10.0	10.0
Classified in a sDCG (%)	4.5	4.5	4.5
Classified in a CNC (%)	3.5	3.5	3.5
Classified in a DME (%)	2.1	2.1	2.1
Classified in an MHC[b] (%)	5.9	5.9	5.9
Classified in a PCG, pDCG, sDCG, CNC, DME and/or MHC (%)	26.2	26.2	26.2

* PCG (pharmaceutical cost groups); pDCG (primary diagnostic cost groups); sDCG (secondary diagnostic cost groups); CNC (costs nursing and care); DME (durable medical equipment); and MHC (multiple-year high-cost groups).

[a] Aggregated to 1,355,008 rows.

[b] Any of the MHC classes except the first two that indicate no consecutive high costs.

Table 4
Comparative evaluation of predictive performance of RE models.

Level	Measure	OLS model	RF	GBM
Individual (n = 5,102,313)	R^2 (%)	32.0	33.5	32.4
	CPM (%)	31.9	32.7	30.0
	MAPE (€)	1984	1961	2029
Subgroup (n = 699,237)[a]	MAPE (€)	1121	1071	1317

[a] The subgroups are analogous to all possible unique combinations of risk classes in the test data.

percentage point improvement on the CPM, and 32 euros on the MAPE. Note that the R^2 is weighted towards cases with extremer expenses [see Equation (10)]. The larger increase in the R^2 compared to the CPM indicates a larger improvement in the prediction for cases with higher expenses. The GBM shows an improvement of 0.4 percentage points in the R^2, but a deterioration in the CPM and MAPE. The deterioration of the CPM and MAPE means that the predictive performance of the GBM is inferior when all cases are weighted equally. Similar to the RF, this also indicates an improvement for cases with high expenses. The superiority of the RF at the individual level translates into an improvement of 50 euros on the MAPE measure for all subgroups compared to the OLS model. The absolute deviation within all conceivable subgroups in the RE model is thus reduced, on average. The GBM performs worse than the OLS regression model in this respect.

Fig. 1 visualizes the net compensation measured by the MER of each RE model per decile of the population ordered by actual expenses. This is to better understand where improvements occur in the distribution of health expenses. The RF outperforms the classic OLS model in all but the first two deciles. The GBM reduces the overcompensation for the lowest deciles significantly but is outperformed in the middle segment of the population. Consistent with the results reported in Table 4, we perceive that the prediction for the upper deciles is better for both the RF and GBM compared to the classic OLS regression model.

4.3. Compensation subgroups based on historical expenses

Fig. 2 reports the MER of all models for subgroups defined using expenses in year *t*-3. As mentioned in Subsection 3.5, the *t*-3 specification filters out incidental expenses, and hence portrays structural outcomes better. Fig. 2 shows that OLS undercompensates the top 15% segment and overcompensates the other two segments. The RF reduces the overcompensation of the broad middle segment of the population from 6 euro to 4 euro, but this comes at the expense of a slight deterioration in the subgroups comprising the top 15% and lowest 15% of expenses. From the individual-level statistics, however, it follows that this reallocation of RE compensations among the three subgroups improves overall predictive accuracy.

Fig. 2 also shows that the GBM drastically reduces the MER of the subgroup with the lowest 15% of expenses from 110 euro to 4 euro. This indicates a significant reduction in the overcompensation of individuals with predictably low expenses. This reduction comes at the cost of the performance in the middle 70% and, to a lesser extent, the highest 15% of year *t*-3 expenses. The individual-level statistics reveal that this reallocation of RE compensations among the three subgroups does not plainly improve total predictive performance of the RE model. In fact, GBM dominates OLS only in terms of R^2.

4.4. Model interpretation metrics

Table 5 reports the marginal effects of the ten most important predictors for the classic OLS model, the RF and the GBM. Overall, the importance of the predictors is comparable across models, while their marginal effects differ. The list of important predictors contains primarily cost-related risk classes. All predictors but two (PCG32 and pDCG14) present in the ranking of the OLS model are also identified in the ranking of the RF or the GBM.

Moreover, three specific findings emerge from the data. First, MHC8, which relates to individuals with three consecutive years of top 0.5% expenditure, ranks highest in relative importance in the RF and GBM, which suggests that this variable is used in a high number of splits in the underlying regression trees. However, the marginal effect (i.e., the main effect) is comparable to the OLS model. The higher relative importance may in part be explained by interactions with other risk classes that are not reflected in the marginal effect. Second, PCG32 and 33, which represent two clusters of drugs with extremely high costs, drop in relative importance in the RF and GBM. The estimated marginal effects are also considerably lower, explaining in part the lower ranking. Finally, sDCGs, which indicate co-morbidity. which effects could be dependent on the pDCG assigned, play a relatively larger role in explaining expenditure in the RF and GBM than in the OLS model.

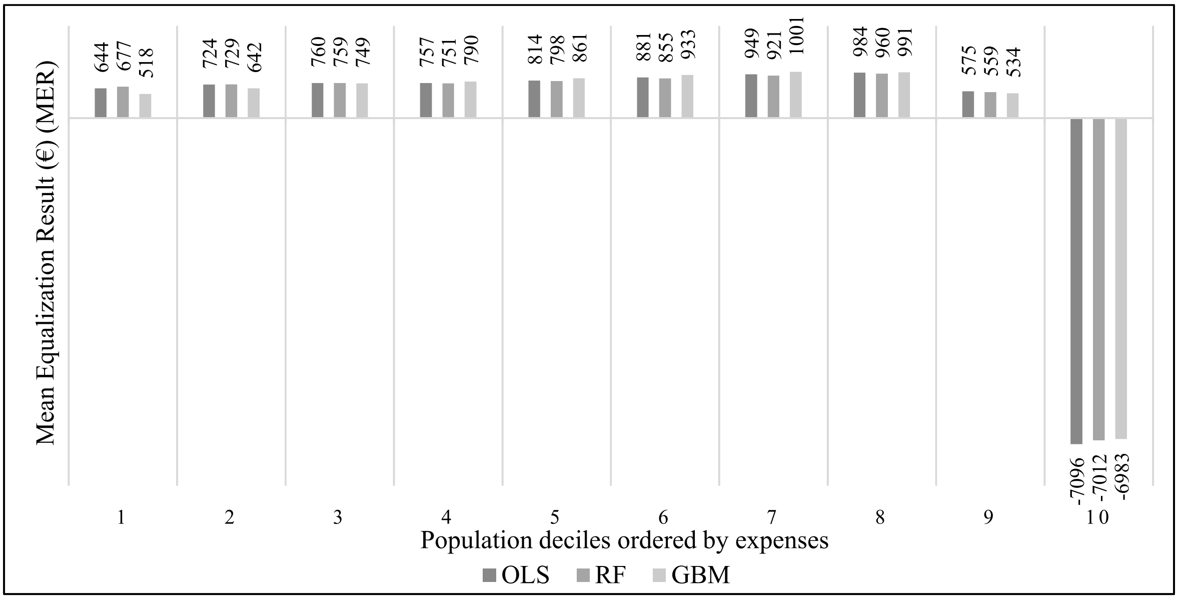

Fig. 1. Mean equalization result (MER) of the population deciles ordered by expenses in year *t*.

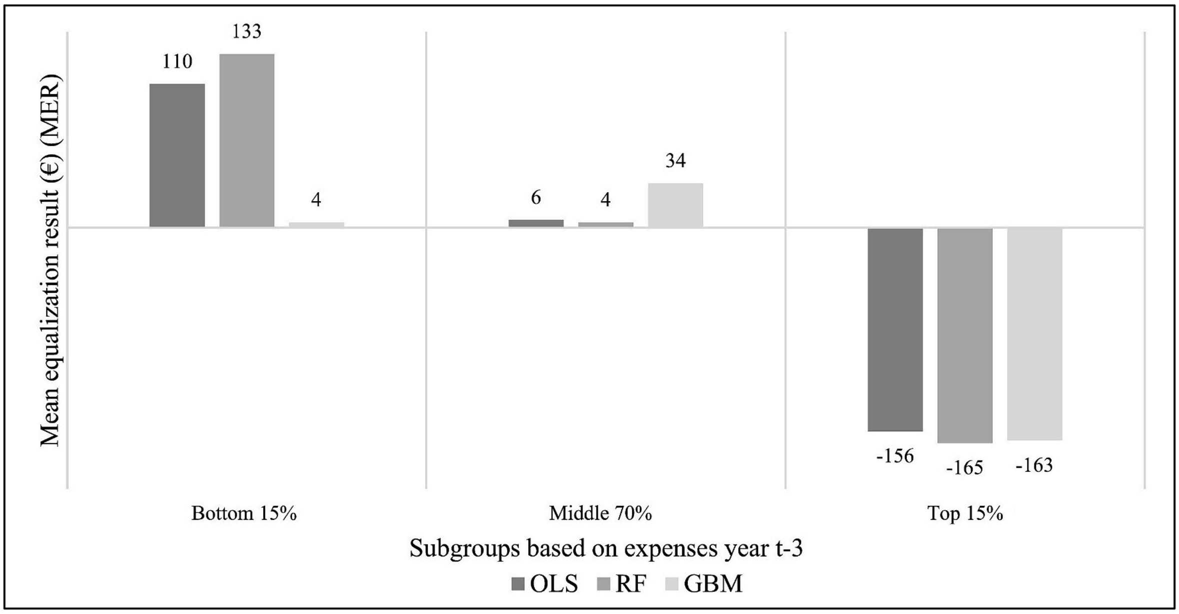

Fig. 2. Mean equalization result (MER) of subgroups based on expenses year *t*-3.

Table 5
Marginal effects of top 10 most important predictors in RF and GBM.

Predictors	OLS		RF		GBM	
	Importance rank[a,b]	Marginal effect	Importance rank[a,b]	Marginal effect[c]	Importance rank[a,b]	Marginal effect[c]
PCG33	1	447,183	4	125,537	4	295,713
CNC7	2	28,687	3	26,334	2	32,256
MHC8	3	43,380	1	48,724	1	48,086
pDCG15	4	51,837	5	49,774	3	72,254
MHC6	5	8891	6	7174	9	8596
MHC7	6	17,235	8	15,004	7	15,966
PCG32	7	193,705	…	…	…	…
CNC4	8	10,178	10	9809	10	12,004
pDCG14	9	64,847	…	…	…	…
PCG29	10	13,182	…	…	5	16,487
PCG0	…	…	2	−1278	…	…
PCG14	…	…	7	1741	…	…
sDCG4	…	…	9	7660	6	10,075
sDCG3	…	…	…	…	8	6271

* PCG (pharmaceutical cost groups); pDCG (primary diagnostic cost groups); sDCG (secondary diagnostic cost groups); CNC (costs nursing and care); DME (durable medical equipment); and MHC (multiple-year high-cost groups).

[a] Based on std. coefficients in case of the OLS model and based on Friedman's method in case of the RF and GBM.

[b] Ordered by the ranking of the OLS model, followed by the RF and GBM.

[c] Can be interpreted similarly to the regression coefficients (i.e., marginal effects) in the classic OLS regression model.

5. Conclusions

Improving RE modeling is critical to reducing adverse selection incentives that may challenge the public goals of essential societal systems like health insurance, education, and lower government service provision. Without effective RE mechanisms, these systems risk becoming exclusive, leading to disparities in societal equity and market efficiency. This study takes the case of regulated health insurance in the Netherlands to demonstrate how regulators can advance RE modeling to improve the accuracy, and consequently effectiveness of RE policies. In that context, our study extends the work of Iommi et al. (2022), McGuire et al. (2021) and Van Veen et al. (2017) by employing population-wide data representative at the national level. This allows us to fully leverage the ability of machine learning to identify unanticipated yet structural groups with meaningful deviation in medical expenditure and provide an authentic benchmark against an established RE model used in practice. More generally, this paper extends the recent stream of research on machine learning applications, particularly, tree-based ensemble methods, in economic analysis (Colak et al., 2022; Gu et al., 2023; Hughes et al., 2022; Kleinberg et al., 2018).

We find that machine learning improves predictive power and the net compensation of selective groups compared to the Dutch RE model based on OLS regression. The RF estimates medical expenditure most accurately, as shown by all model fit metrics, with improvements particularly for groups with moderate to high actual medical costs. The GBM selectively improves prediction for groups with the lowest or highest actual costs. Note that the actual costs include, to an extent, incidental expenses that are typically not in scope for compensation. The subgroup results based on historical expenses (year *t-3*) are less sensitive to incidental expenses and add insight in the mitigation of structural adverse selection incentives. The overall better model fit of the RF translates into an improvement over the classic OLS model in the net compensation of the large midsegment with moderate historical expenses, while results in the other subgroups are similar. The GBM notably reduces the overcompensation of the subgroup with the lowest 15% of historical expenses, which concerns enrollees that are especially prone to risk selection. The GBM could offer opportunities to mitigate these incentives, while combining both algorithms may yield the best overall results.

The innovation machine learning algorithms offer over statistical models is their capacity to capture complex interaction effects without a priori assumptions. Since all explanatory variables concern binary risk classes and are kept constant across models, the observed improvements can be attributed to interaction effects detected by the RF and GBM, which were omitted in the OLS model. For the RF, this implies that unforeseen risk factor interactions account for an additional 1.5% of the variance in medical expenditure, translating to approximately 571 million euros across all individuals in the Dutch context (see Table 4). In the case of the GBM, the additional 0.4% variance explained still corresponds to 152 million euros. These improvements suggest that the OLS model may be improved considerably by solely introducing new interaction effects. Delving into the outcomes of the RF and GBM, it is also worth noting where they differ. The RF leverages bootstrap aggregation, where the estimates of multiple trees trained on different subsets of data are combined to come to an overall more robust prediction on unseen data (Breiman, 2001a). The GBM instead uses boosting to iteratively improve the prediction of each regression tree (Friedman, 2001). This can lead to a focus on where residuals are largest and could explain why improvements are primarily perceived at the tails of the distribution of expenses.

From a conceptual perspective, we observe that the machine learning methods place greater emphasis on the relationship of other risk classes to medical expenditure. Notably, the multiple-year high-cost group which relates to individuals with three consecutive years of top 0.5% expenditure, plays a pivotal role in explaining expenditure in the RF and GBM. The estimated main effect is however comparable to the OLS model, suggesting that its importance in part stems from interactions with other risk classes. Furthermore, it becomes evident that the top pharmaceutical cost groups, relating to clusters of drugs with extremely high costs, are given less weight in the RF and GBM, and have a considerably smaller impact on medical expenditure. Finally, we observe that secondary diagnostic cost groups (sDCGs), which were introduced to account for co-morbidity, are given more weight in the RF and GBM. Despite it being foreseeable that the impact of sDCGs on medical expenditure may be dependent on the primary DCG assigned, the state-of-the-art OLS model does not allow for interactions between sDCGs and pDCGs. In the RF and GBM, the sDCGs are allowed to interact with other risk classes, which may explain their importance in these models.

A few considerations must be made regarding this study. Our analyses were performed primarily from the standpoint of the health regulator, who wishes to minimize risk selection by private health insurers (promoting access equity) and to level the playing field among insurers (increasing market efficiency). Since RE is executed at the population level in the Netherlands, we refrained from modeling approaches that counter this by, e.g., segmenting the data and creating separate models per segment. Similarly, this study foregoes opportunities to introduce new risk classes or to improve their specification to ensure a close comparison to the existing RE model in the Netherlands. Note, however, that such approaches could potentially further optimize predictive performance, and are hence worth considering in future studies (McGuire et al., 2021). Our findings thus primarily relate to insurance regulators. The extent to which an individual insurer may benefit from machine learning will depend on the maturity and quantity of their own data.

Furthermore, an oft-expressed critique of machine learning is that the absence of an explicit equation compromises interpretability and can hinder implementation in policy context. Here, we present three counterarguments. First, the interpretability challenge is not dissimilar to that encountered by complex econometric models. For both holds that this challenge can be addressed through functional analysis. The interpretable machine learning (IML) research community is continuously working on addressing issues of interpretability through new methods. In this study, we leverage some of these methods to extract important main effects from the RF and GBM. For further reference, Chan and Mátyás (2022) provide more examples of IML methods to address transparency in economic analysis. Second, transparency varies among machine learning algorithms, with tree-based methods like RFs excelling in this regard. Hence, in policy settings where transparency is important, these methods may be one of the more suitable options. The final counterargument is specific to RE context, where the lack of an explicit equation can also have advantages. For instance, compromising transparency could restrict efforts from insurers to 'game the system' through up-coding and other risk-inflating behaviors (Geruso and Layton, 2020).

The insights from this study may have several applications for regulators seeking to improve their RE implementation. Our findings relate to RE in health insurance markets but could also be relevant for other healthcare financing applications, such as risk adjustment of healthcare provider payments and regional funds, and for other economic settings where RE-like policies are common, such as local services and education. Common feasibility criteria within RE schemes, which include transparency and practicality, suggest using a combination of classic statistics and machine learning modeling methods, combining the pros of both types of methods – i.e., the higher transparency of classic regression and the higher predictive accuracy of machine learning. Hence, it is foreseeable that machine learning methods find their application as a complement to OLS, rather than as a replacement. Van Veen et al. (2017) and McGuire et al. (2021) provide frameworks for applying machine learning for the identification of interactions among risk adjusters and variable selection. Applying machine learning as a complement to existing classic linear regression models could allow regulators to leverage the former's advantages without foregoing feasibility criteria of practicality and transparency.

The superior performance of machine learning over the state-of-the-art Dutch RE model also highlights a potential vulnerability that insurers could exploit. While our study adheres to common substantive criteria in RE, utilizing predefined risk classes and not employing more recent data than typically available to regulators, we believe that the machine learning models could yield even better results if these constraints were relaxed. Notably, insurers are not bound by these criteria and face fewer barriers to adopting machine learning for selection strategies. While Dutch law prohibits denial of access and premium differentiation for basic insurance, insurers still have alternative tools for risk selection. These encompass product design, service offerings, and marketing strategies, such as selection via supplementary insurance variations, targeting specific demographic groups like highly educated individuals, and segmenting their insurance products and group arrangements (Van de Ven et al., 2017, 2023).

In conclusion, we find that the studied machine learning methods hold promise for regulators to improve RE in domains in society where adverse selection reduces access equity and harms market efficiency, and therefore recommend them to be embedded into the maintenance of regulatory RE models. In consideration of transparency, we advocate for the initial use of machine learning to inform conventional classic linear regression modeling. To facilitate this, further research is recommended on how to best translate parameters in machine learning models to existing RE models. Finally, we recognize the potential and increasing threat that lies in the use of machine learning for risk selection purposes by agents, and therefore urge regulators to invest in further research on vulnerabilities of current RE models that may be exposed through the application of machine learning. Importantly, adopting machine learning for the improvement of RE should help regulators to reduce such threats in advance and to identify them when they materialize, instead of being reactive.

Declaration of competing interest

None.

Data availability

The authors do not have permission to share data.

Acknowledgements

We would like to gratefully thank the editor and the two anonymous reviewers for their considered comments that helped improve this article. We would like to also extend our gratitude to Johan Visser and Professor Pierre Koning for their valuable feedback during this study. Finally, we would like to thank the Ministry of Health, Welfare and Sport and the Dutch Association of Health Insurers for providing access to the administrative dataset used in this study.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Aguirre, I., Beitia, A., 2017. Modelling countervailing incentives in adverse selection models: a synthesis. Econ. Modell. 62 https://doi.org/10.1016/j.econmod.2017.01.007.

Alanis, E., 2022. Forecasting betas with random forests. Appl. Econ. Lett. 29 (12) https://doi.org/10.1080/13504851.2021.1912278.

Bos, F., 2013. Economic theory and four centuries of fiscal decentralisation in The Netherlands. OECD J. Budg. 12 (2) https://doi.org/10.1787/budget-12-5k8zpd5cczd8.

Breiman, L., 2001a. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., 2001b. Statistical modeling: the two cultures. Stat. Sci. 16 (3), 199–231. https://doi.org/10.2307/2676681.

Brown, S., Hole, A.R., Kilic, D., 2014. Out-of-pocket health care expenditure in Turkey: analysis of the 2003–2008 household budget surveys. Econ. Modell. 41, 211–218. https://doi.org/10.1016/J.ECONMOD.2014.05.012.

Carmona, P., Climent, F., Momparler, A., 2019. Predicting failure in the U.S. banking sector: an extreme gradient boosting approach. Int. Rev. Econ. Finance 61. https://doi.org/10.1016/j.iref.2018.03.008.

Chan, F., Mátyás, L. (Eds.), 2022. Econometrics with Machine Learning, vol. 53. https://doi.org/10.1007/978-3-031-15149-1.

Colak, G., Fu, M., Hasan, I., 2022. On modeling IPO failure risk. Econ. Modell. 109 https://doi.org/10.1016/j.econmod.2022.105790.

Cui, K., Li, B., Wang, H., 2021. Quantitative analysis of health insurance reform in China: pure consolidation or universal health insurance? Econ. Modell. 101, 105550 https://doi.org/10.1016/J.ECONMOD.2021.105550.

Du, Y., 2023. Health investment and medical risk: new explanations of the portfolio puzzle. Econ. Modell. 127, 106442 https://doi.org/10.1016/j.econmod.2023.106442.

Eichhorst, A., 2007. Evaluating the need assessment in fiscal equalization schemes at the local government level. J. Soc. Econ. 36 (5) https://doi.org/10.1016/j.socec.2007.01.009.

Ellis, R.P., McGuire, T.G., 2007. Predictability and predictiveness in health care spending. J. Health Econ. https://doi.org/10.1016/j.jhealeco.2006.06.004.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

Geruso, M., Layton, T., 2020. Upcoding: evidence from medicare on squishy risk adjustment. J. Polit. Econ. 128 (3) https://doi.org/10.1086/704756.

Geruso, M., Mcguire, T.G., 2016. Tradeoffs in the design of health plan payment systems: fit, power and balance. J. Health Econ. 47, 1–19. https://doi.org/10.1016/j.jhealeco.2016.01.007.

Gu, Z., Li, Y., Zhang, M., Liu, Y., 2023. Modelling economic losses from earthquakes using regression forests: application to parametric insurance. Econ. Modell. 125, 106350 https://doi.org/10.1016/j.econmod.2023.106350.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning. Springer 18 (4), 746. https://doi.org/10.1007/b94608, 2001.

Hocking, R.R., 2003. Methods and Applications of Linear Models: Regression and the Analysis of Variance. Wiley.

Hughes, N., Soh, W.Y., Lawson, K., Lu, M., 2022. Improving the performance of micro-simulation models with machine learning: the case of Australian farms. Econ. Modell. 115 https://doi.org/10.1016/j.econmod.2022.105957.

Iommi, M., Bergquist, S., Fiorentini, G., Paolucci, F., 2022. Comparing risk adjustment estimation methods under data availability constraints. Health Econ. 31 (7), 1368–1380. https://doi.org/10.1002/hec.4512.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S., 2018. Human decisions and machine predictions. Q. J. Econ. 133 (1) https://doi.org/10.1093/qje/qjx032.

Ladd, H.F., Fiske, E.B., 2011. Weighted student funding in The Netherlands: a model for the U.S. J. Pol. Anal. Manag. 30 (3) https://doi.org/10.1002/pam.20589.

Layton, T.J., Ellis, R.P., McGuire, T.G., van Kleef, R., 2017. Measuring efficiency of health plan payment systems in managed competition health insurance markets. J. Health Econ. 56 https://doi.org/10.1016/j.jhealeco.2017.05.004.

McGuire, T.G., Schillo, S., van Kleef, R.C., 2020. Very high and low residual spenders in private health insurance markets: Germany, The Netherlands and the U.S. Marketplaces. Eur. J. Health Econ. 22 (1) https://doi.org/10.1007/s10198-020-01227-3.

McGuire, T.G., Zink, A.L., Rose, S., 2021. Improving the performance of risk adjustment systems constrained regressions, reinsurance, and variable selection. Am. J. Health Econ. 7 (4) https://doi.org/10.1086/716199.

Murthy, V.N.R., Okunade, A.A., 2016. Determinants of U.S. health expenditure: evidence from autoregressive distributed lag (ARDL) approach to cointegration. Econ. Modell. 59, 67–73. https://doi.org/10.1016/J.ECONMOD.2016.07.001.

Nicoletti, C., Best, N., 2012. Quantile regression with aggregated data. Econ. Lett. 117 (2) https://doi.org/10.1016/j.econlet.2012.06.011.

Oliva, M., Carles, M., 2013. Price dispersion in the private health insurance industry: the case of Catalonia. Econ. Modell. 31 (1), 177–182. https://doi.org/10.1016/J.ECONMOD.2012.11.029.

Oskam, M., van Kleef, R.C., van Vliet, R.C.J.A., 2023. Improving diagnosis-based cost groups in the Dutch risk equalization model: the effects of a new clustering method and allowing for multimorbidity. Int. J. Health Econ. Manag. https://doi.org/10.1007/s10754-023-09345-0.

Ozden, E., Guleryuz, D., 2022. Optimized machine learning algorithms for investigating the relationship between economic development and human capital. Comput. Econ. 60 (1) https://doi.org/10.1007/s10614-021-10194-7.

Roza, M., Hagan, K., Anderson, L., 2021. Variation is the norm: a landscape analysis of weighted student funding implementation. Public Budg. Finance 41 (1). https://doi.org/10.1111/pbaf.12276.

Schokkaert, E., Dhaene, G., Van De Voorde, C., 1998. Risk Adjustment and the Trade-Off between Efficiency and Risk Selection: an Application of the Theory of Fair Compensation. Health Economics. https://doi.org/10.1002/(SICI)1099-1050(199808)7:5<465::AID-HEC365>3.0.CO;2-9.

Siegel, A.F., 2016. Chapter 12 - multiple regression: predicting one variable from several others. In: Practical Business Statistics, seventh ed.

Van de Ven, W.P.M.M., Ellis, R.P., 2000. Chapter 14 Risk adjustment in competitive health plan markets. In: Handbook of Health Economics, vol. 1. Issue PART A, pp. 755–845. https://doi.org/10.1016/S1574-0064(00)80173-0.

Van de Ven, W.P.M.M., Hamstra, G., van Kleef, R., Reuser, M., Stam, P., 2023. The goal of risk equalization in regulated competitive health insurance markets. Eur. J. Health Econ. 24 (1), 111–123. https://doi.org/10.1007/S10198-022-01457-7/METRICS.

Van de Ven, W.P.M.M., Van Vliet, R.C.J.A., Van Kleef, R.C., 2017. How can the regulator show evidence of (no) risk selection in health insurance markets? Conceptual framework and empirical evidence. Eur. J. Health Econ. 18 (2), 167–180. https://doi.org/10.1007/S10198-016-0764-7.

Van Veen, S.H.C.M., Van Kleef, R.C., Van de Ven, W.P.M.M., Van Vliet, R.C.J.A., 2017. Exploring the predictive power of interaction terms in a sophisticated risk equalization model using regression trees. Health Econ. 27 (2), 1–12.

Withagen-Koster, A.A., Van Kleef, R.C., Eijkenaar, F., 2018. Examining unpriced risk heterogeneity in the Dutch health insurance market. Eur. J. Health Econ. 19 (9) https://doi.org/10.1007/s10198-018-0979-x.

Zhang, Y., Yang, Y., 2015. Cross-validation for selecting a model selection procedure. J. Econom. 187 (1) https://doi.org/10.1016/j.jeconom.2015.02.006.