Identifying under- and overcompensated groups in the Dutch Risk Equalization Model using stacking as a benchmark-model.

Erasmus University Rotterdam - Erasmus School of Economics
Master Thesis: Quantitative Finance

Author:
Name student: Thijmen de Vries
Student ID number: 456707

Supervision:
Supervisor ESE: Dr. (Andreas) A. Pick
Second assessor: Dr. (Marina) M. Khismatullina
Supervisor PwC: Dr. (Suzanne) S.H.C.M. van Veen

Date final version: October 10, 2022

Abstract

In the Netherlands, a risk equalization system provides health insurers with an ex-ante compensation for predictable differences in somatic health care expenditure made by individuals. This ex-ante compensation is estimated by a predictive model using OLS. However, this model does not produce accurate results and therefore insurers are under- or overcompensated compared to the actual incurred somatic health care expenditure, for specific groups of insured. In this research, data regarding the full Dutch population (N=16,327,282) is used to benchmark the predictive performance of the 2022 Dutch risk equalization model against a set of alternative models: a Ridge regression, Ordered Logit regression, Random Forest model, Gradient Boosted model, and a stacked model. The latter combines the aforementioned alternative models in one predictive model. Predictive performance of all evaluated models is assessed on subgroup level. These subgroups are created based on the result on risk equalization over the years 2019 to 2021. It is found that the Random Forest, Gradient Boosted and stacked model outperform the current OLS risk equalization model. The predictive outcomes of these models compared to the OLS model indicate subgroups that are potentially subject to risk selection, as these subgroups are significantly under- or overcompensated by the ex-ante compensation estimated by the OLS model.

Acknowledgements

I hereby thank the Dutch Ministry of Health, Welfare and Sport and ZN, an assocation of health insurers in the Netherlands, for their permission to use the data to conduct this research..

I would also like to thank my supervisor from the Erasmus School of Economics, dr. A. Pick, for his time and valuable comments during the period of this research. I also thank dr. M. Khismatullina for assessing this thesis.

Furthermore, I would like to thank my colleagues at PwC for everything they taught me and all moments that we already shared together. A special word of thanks to dr. S.H.C.M. Van Veen for introducing me to the research field of risk equalization and for her time and valuable feedback during the period of conducting this research. I also want to thank dr. W. Karman for his guidance and challenging feedback during my whole period as a working student at PwC.

A final word of thanks goes out to my family and friends, who have supported me throughout my entire study period.

Table of Contents

1 Introduction **5**

2 Data **11**

2.1 Data handling 11

2.2 Variables used in the Dutch risk equalization 12

2.3 Somatic health care expenditure in the Netherlands 14

2.3.1 Actual somatic health care expenditure in 2016-2019 15

2.3.2 Annualized somatic health care expenditure in 2019 16

2.4 Subgroups defined to evaluate model predictive performance 17

3 Methodology **19**

3.1 Current risk equalization method: linear regression 19

3.2 Regularized regression model 20

3.3 Conditional Density Approximation using Ordered Logit 21

3.4 Regression Tree models 22

3.4.1 Random Forest 23

3.4.2 Gradient Boosted Regression Tree 24

3.5 Combined model: stacked generalization 25

3.5.1 Optimal model specifications of the underlying models 25

3.5.2 Combined model using stacked generalization 26

3.6 Evaluation of predictive performance 27

3.6.1 Individual level evaluation metrics: R^2, CPM and MSE 28

3.6.2 Subgroup level evaluation metric: MPE 29

4 Results **30**

4.1 Regularized regression model 30

4.2 Ordered Logit model 31

4.3 Random Forest 32

4.4 Gradient Boosted model 33

4.5 Stacked predictive model 34

4.6 Evaluation of predictive performance 34

4.6.1 Individual level predictive performance 35

4.6.2 Subgroup level predictive performance 35

5 Discussion and conclusion **39**

5.1 Discussion 39

5.2 Conclusion 40
5.3 Recommendations for further research 41

References **42**

6 Appendix **45**
6.1 Variables in the Dutch risk equalization models over time (2019-2022). 45
6.2 Descriptive statistics for actual somatic health care expenditure in the full data set. 46
6.3 Descriptive statistics for annualized somatic health care expenditure in 2019. 47
6.4 Descriptive statistics for the subsets of data used in this research. 49
6.5 Predictive results: unscaled expenditure level 50
6.6 Final model estimation results. 51
6.6.1 Model estimation results: parametric regressions 51
6.6.2 Model estimation results: tree-based regression models 58

1 Introduction

With the introduction of the Health Care Law in the Netherlands in 2006, the Dutch government set a duty of acceptance of people and prohibited differentiation in premiums charged by health insurers (Stam et al., 2020; Zorgautoriteit, 2019). This still existent ban on premium differentiation creates specific groups of insured (Stam et al., 2020) which are predicted to be profitable or loss-giving for health insurers. This leads to a financial incentive for insurers to target specific groups in the Dutch population (Zorgautoriteit, 2019) for insurance policies.

To counter this incentive, the risk equalization system is created. The Dutch risk equalization system is a regulated scheme which provides health insurers with an annual ex-ante compensation for the predictable differences in health care expenditure per individual (Van Kleef et al., 2012). The goal of this system is to create a level playing field between health insurers, to reduce the incentive for risk selection among health insurers and to stimulate the efficiency of the Dutch health care (Van Veen et al., 2015). The ex-ante compensation is based on a prediction model, which is reviewed and improved on a constant basis. However, as the risk equalization model does not predict adequately, insurers are still significantly under- or overcompensated for specific groups in the Dutch population (Van Kleef et al., 2017).

When the characteristics of these groups are known, insurers can take actions thereon, for example in terms of marketing or policy creation to target the overcompensated people (Van Kleef et al., 2017; Zorgautoriteit, 2019). This targeting competition leaves specific insurers with portfolios over-represented with undercompensated people (Stam et al., 2020), leading to higher premiums asked by these insurers which scares off healthy people in their portfolio. This adverse selection disrupts the health insurance market, as it is an unequal level playing field between insurers. Premiums are not based on efficiency and quality, but a reflection of the insurer's portfolio (Van Veen, 2016). Despite the risk equalization system, the incentive to target a specific group of profitable people therefore remains. There are concrete signals for risk selection, as the focus of Dutch health insurers is to create policies with restrictive conditions to target healthy people (Zorgautoriteit, 2021). This risk selection is a threat to the goals of the Dutch health care system as intended in 2006 (Van Kleef et al., 2020; Visser et al., 2014).

In this research, the current magnitude and nature of the incentive for risk selection is investigated further with the use of a not earlier applied technique on the Dutch risk equalization data. The current Dutch risk equalization model uses linear regression to estimate the ex-ante compensation, but research has proven that more advanced estimation techniques produce more accurate predictions, e.g. Stam et al. (2020), Ellis et al. (2018) and Jones et al. (2015). These techniques are however not used in practice yet, because they are hard to implement and lack transparency and interpretability (Ellis et al., 2018). This is regarded as highly important in the risk equalization. Given this disinclined attitude on replacing the current estimation technique with more advanced models, this research aims to evaluate the current technique with insights gained from multiple dis-

similar algorithms, combined into a combination of models (Bates & Granger, 1969; Van Der Laan et al., 2007; Wolpert, 1992). This method of combining models into one is referred to as stacked generalization.

The model created in this research serves as a policy evaluation instrument for the Dutch risk equalization system. It evaluates the predictive power of the current Dutch risk equalization model against a more advanced method. The aim of this research is to discover subgroups in the Dutch population for which an advanced model provides significantly better predictions compared to the current linear model. The advanced model provides an insight in the information that a Dutch health insurer could have, benchmarked against the actual risk equalization method. It is tested, whether an insurer can identify significantly under- or overcompensated subgroups based on this information. It therefore provides an insight in the possibilities for risk selection from the perspective of a health insurer.

Van Veen et al. (2015) indicate that this analysis on incentives for risk selection should be done on non-random subgroup level. The results for other evaluation levels, such as the full sample, are likely to be the cleared effect of under- and overcompensations and therefore lead to contaminated conclusions. The evaluation of models is therefore based on subgroup level, which are defined prior to the model estimation. The central question of this research is:

Which added value brings the application of the stacked algorithm to serve as a policy evaluation instrument for the current Dutch risk equalization model?

This question can be divided in the following sub-questions:

- In which way can a stacked model be developed as a benchmark model? What conditions does a policy evaluation benchmark have to meet?
- What is the predictive power of this benchmark model? (in terms of individual and subgroup level)
- What is the performance of the current risk equalization model compared to the benchmark model? (in terms of individual and subgroup level)
- What are the characteristics of the identified groups for which the risk equalization system is a less good predictor for health care costs compared to the benchmark model?

The current methodology used in the Dutch risk equalization system is OLS. The use of this linear regression comes with several assumptions on the disturbances, such as homoskedasticity in the covariance matrix and a joint normal distribution (Heij et al., 2004). Given the distribution of health care expenditure in a population, these assumptions might be too stringent. Researches indicate that health care expenditure data incorporates uncommon features, such as non-negative

values and many observations with 0 expenditure within a year (Jones et al., 2015; Vimont et al., 2022).

Jones et al. (2015), using an English health care expenditure data set for the period 2007-2008, also provide evidence against the assumptions of homoskedasticity and a normal distribution of the disturbances. They respectively found a trend between the mean predicted health costs in pre-specified quantiles and the variance of these costs, and non-normal values for the higher moments in the distribution (skewness, kurtosis). Presented with these findings literature advises to use other, more flexible approaches to model the complex distribution of health care data. These more flexible methods are researched widely, both as an addition to the linear regression estimation method as well as to replace this method.

A major advantage of more flexible model specifications is feasibility to utilize the data structure in a more effective way, with extended possibilities to incorporate interaction terms between explanatory variables (Rose, 2016; Stam et al., 2020). Such a flexible model can also be used as an addition to the current linear risk equalization method. Van Veen et al. (2017) uses regression trees to research interactions between variables in the Dutch risk equalization system. Those interaction terms that capture variation in the health care expenditures unexplained by the current model, are added to the to the risk equalization model by Van Veen et al. The resulting model including the selected interaction terms shows increased predictive performance compared to the standard model. This serves as an example on how to improve the current model with insights gained from alternative models.

In this thesis, a set of more flexible model specifications is tested on the Dutch health care data and eventually combined into one predictive algorithm, as proposed by Bates and Granger (1969) in their research on a combination of forecasts. They use a combination of two forecasts and conclude that the Mean Squared Error of such a combined model is lower compared to the separate forecast models. Later research by Wolpert (1992) generalizes this idea to incorporate as many algorithms as intended in the combined forecast. This method is called stacked generalization.

The use of a combined predictive model is attractive because it can combine several algorithms that learn from the underlying data in different ways (Bates & Granger, 1969; Van Der Laan et al., 2007; Wolpert, 1992). The combination of models therefore overcomes the choice of model specification, a task specifically complex in risk equalization as several models have established divergent results in earlier literature in terms of predictive power, both on individual level and on subgroup level.

Stacked generalization involves two steps. At first, different *base* models are estimated to predict the dependent variable. In the next step, these models are combined into one final prediction using the base model predictions made on a hold-out validation set or using k-fold cross validation. These base model predictions are the explanatory variables in the final model.

The use of k-fold cross validated predictions as explanatory variables in a stacked model is called a Super Learner model (Van Der Laan et al., 2007). Such a Super Learner is applied earlier

to American health expenditure data by Rose (2016). The set of base models tested in this research involves parametric regression models, penalized regression models (Lasso, Ridge and Elastic Net) and machine learning models such as Decision Trees, Random Forests and Neural Networks. The algorithms are trained in two ways: on the full set of descriptive variables used in the risk equalization model in the United States and on a subset of these variables, selected by means of a Random Forest which selects the top 10 variables with highest predictive importance.

All models trained on the subset of variables yielded lower predictive power compared to the full models. However, the impact of the variable reduction was relatively low. This is promising, as covariate reduction in a model leads to higher interpretability of a model, an important feature in risk equalization.

In both the full and more parsimonious model specifications estimated by Rose, the Super Learner performs better than every single algorithm in terms of predictive performance (R^2) on individual level, providing recent evidence for the claim that a combination of forecasts improves the predictive performance (Bates & Granger, 1969). This result serves as motivation for further research on the use of combined predictive models, as done in this thesis.

Stam et al. (2020) compare the use of a Random Forest and a Gradient Boosted model to the actual OLS model for somatic health care expenditure of the total Dutch population in 2018. On individual level the models are evaluated using R^2, `CPM` and `GGAA`[1]. The Random Forest outperforms the OLS model in all these metrics, the Gradient Boosted Machine fails to outperform in terms of `CPM` and `GGAA`. However, the results for all models are very close to each other.

On subgroup level the three compared models present more divergent results. The three subgroups evaluated for this research are based on the level of health care expenditure in 2015 (15% lowest spending on health care in 2015, 70% middle class and 15% highest spending in 2015). Stam et al. (2020) indicate that the OLS model overcompensates the first two groups and undercompensates the last group. The Random Forest model reduces the overcompensation of the middle class but increases the under- and overcompensation of the other two subgroups. As the two subgroups with lowest and highest spending on health care are specifically subject to respectively positive and negative risk selection by health insurers, the Random Forest does not mitigate this incentive to a large extent. In comparison, despite not having a better predictive power on the individual level compared to OLS, the Gradient Boosted model presents interesting results on subgroup level. For the group with 15% lowest health care expenditure in 2015 it firmly reduces the overcompensation compared to OLS, the researchers find.

Stam et al. (2020) indicate that this currently overcompensated group is the main target for positive risk selection by Dutch health insurers and thus this result is interesting, as it takes away part of the incentive for positive risk selection. In contrast, the over- and undercompensation for the other two evaluated groups increase with the use of the Gradient Boosted model compared to

[1] R^2 and CPM are used in this thesis and discussed in detail in section 3.6.1. GGAA measures the weighted mean absolute prediction error.

OLS. Given these divergent results on individual and subgroup level, there is no clear best option among the two alternative algorithms compared to OLS. Therefore, the stacked combination of predictive models used in this thesis is attractive as it can incorporate both as a base model instead of choosing between them.

Using a French health care spendings dataset, research on a more detailed subgroup level is presented (Vimont et al., 2022). The researchers test the use of advanced methods, which can indicate complex interactions and relations between variables used. Using both a Generalized Linear Model, Random Forest, and a Neural Network they find that all these models overestimate the health care expenditure for the subgroup with lowest observed expenditure (< 100) in a year. For the high-expenditure subgroups (> 5000 and > 15000) the Random Forest model is the most accurate. It yields a better performance in the tail of the distribution as the algorithm is more robust to extreme values in the data, the researchers conclude (Vimont et al., 2022).

Next to researched machine learning methods, parametric regression techniques are evaluated for risk equalization purposes as well. Jones et al. (2015) tested the performance of sixteen parametric and semi-parametric regression methods, where each method makes different assumptions on the distribution of health care expenditure. The set of models involves both widely used and more uncommon methods. Widely used models such as linear regression, log-transformed linear regression, and log-link gamma-variance generalized linear regression perform amongst the worst methods in terms of Mean Prediction Error, Mean Absolute Prediction Error, and Root Mean Squared Error. The approximation of the conditional density function as first proposed by Gilleskie and Mroz (2004) yields the most promising results as indicated by Jones et al. (2015). In this method, the continuous health care expenditure is divided into intervals. The conditional probability for an insuree to fall within an interval given the set of explanatory variables is estimated via both Multinomial and Ordered Logit regressions. The resulting probabilities are then multiplied by the mean health care expenditure of each created interval to obtain predictions for health care expenditure. A similar approach to model health care expenditure is followed in one of the models of this thesis.

In addition, research performed by Van Kleef et al. (2015) focuses on the application of constrained regression on the Dutch health care expenditure data. This method is potentially attractive, as it can a-priori specify certain groups for which the under- or overcompensation in the risk equalization system is reduced to a pre-specified number. Van Kleef et al. (2015) assess multiple variations of subgroups in constrained regression which present similar results. The predictive results increase for the subgroups which a constrain. This however comes at the cost of deteriorated results for groups incorporated in the traditional risk equalization by means of a variable, the researchers conclude. The use of constrained regression for risk equalization therefore should be motivated by the consideration between these two effects.

In this consideration, the magnitude of these effects plays an important role. Van Barneveld et al. (2000) argue that small under- and overcompensations for groups of insured are not disruptive for

the equalizing effect of the system, as risk selection by means of targeting subgroups comes at a cost as well. Next to this, Ellis and McGuire (2007) state that not only the magnitude and predictability of under- and overcompensations are important, the predictability of the feature or subgroup that is under- or overcompensated for matters as well. An undercompensation of people with highly predictable health care expenditure is therefore more serious than an undercompensation of people with less predictable expenditure patterns.

In the literature there is a clear industry standard in terms of model evaluation. Metrics often used in papers are the R^2, `CPM` (Cumming's Prediction Measure), `MPE` (Mean Prediction Error), and `MSE` (Mean Squared Error). Van Veen (2015) provides an extensive overview of measures used in risk equalization literature, which incorporates the abovementioned metrics as well. These metrics indicate the fit of the model to the data and therefore present meaningful insights. However, a higher predictive performance in terms of these metrics evaluated on individual level does not necessarily guarantee a model to outperform other models in the mitigation of risk selection incentives on subgroup level. It is therefore of high interest to also evaluate the models on pre-defined, selective subgroups that differ from subgroups used in the model estimation, as a group-level evaluation indicates incentives for risk selection (Gupta, 2020; Van Veen et al., 2015; Van Veen, 2016).

None of the above-mentioned metrics is designed for a subgroup analysis. They can be applied to subgroups, as is done by Jones et al. (2015) in the evaluation of the `MPE` per subgroup in the data set. This however is not done widely. This is a discrepancy between the commonly used evaluation metrics and the goal of risk equalization models (Vimont et al., 2022), which is the mitigation of risk selection incentives. Therefore, in this research a subgroup level evaluation of the models is incorporated. For a detailed discussion of the evaluation metrics used in this research, see section 3.6.

2 Data

To answer the research question stated in the introduction, health care expenditure data from the Netherlands for the period 2016-2019 is used. Permission to use this data is granted by the Dutch Ministry of Health, Welfare and Sport and ZN, an association of health insurers in the Netherlands. This data is used in practice to estimate the Dutch risk equalization models for 2019 to 2022. The Dutch risk equalization model corresponding to year t is estimated with data from year *t-3* as this is the most recent data for which information is available in complete and definite form. The observed health care expenditure in *t-3* is made representative for year t, such that the cost level in the data is equal to the cost level in year t[2]. In this chapter, at first data handling is discussed. Next the variables used in this research are described in detail and descriptive statistics are discussed. At last, the subgroups used for evaluation are created.

2.1 Data handling

The observations in the data correspond to pseudonymized individuals in the Netherlands. It is not necessarily the case that each observation corresponds to a unique pseudonymized individual, because each year there is a group of insurees who change their health insurer during the year. These people are present twice in the dataset for the year in which they changed insurer, as part of the insurees' year corresponds to one health insurer and a part corresponds to another. These observations are summed for this research. This leaves one observation for each unique individual.

Next to this, in each year there are new-borns, people who pass away, immigrants, and emigrants. This group of people is present in the dataset, but they only have information corresponding to the period that they were Dutch inhabitants. For example, if someone passes away on the 1st of July in 2019, they have only been present in the Netherlands for 6 months of that year. Their health care expenditure information therefore only corresponds to that period as well. To recognize this, a variable which denotes the number of days that is accounted for in this observation, is present in the data. This variable is transformed into a weight per observation, by dividing it through the total number of days in each year.

The actual health care expenditure for such an individual who is only partly present in the Netherlands in a calendar year is divided by this weight. Dividing each observation in the data by the weight of that observation results in annualized weighted health care expenditure and observations that are in insured years. An insured year is equal to a full calendar year, the example individual who passes away on the 1st of July corresponds to 0.5 insured years. If this person has already accumulated €100 of health care expenditure in that period, this is divided by a weight of 0.5. This results in an annualized expenditure of €200. These annualized expenditures are used for further

[2]This is done prior to receiving the data. Data is received in the form such that 2019 expenditure is scaled to match the 2022 cost level.

analysis and estimation of the risk equalization model. The descriptive statistics, discussed in section 2.3.1, are based on the actual (unweighted) health care expenditures to present actual costs rather than annualized (weighted) cost information. Annualized expenditure for 2019 is discussed in section 2.3.2.

2.2 Variables used in the Dutch risk equalization

The data used in this research contains individual-level information for all Dutch insured in the period 2016-2019. Each individual is pseudonymized, such that data cannot be traced back to specific people. For each observation, both financial information and all variables used in the Dutch risk equalization system are present.

The variables used to estimate the Dutch risk equalization model are either morbidity-based variables or variables with demographic information per individual. Together, these variables determine the defined health status of an individual for the risk equalization model. The 2022 risk equalization model consists of 13 variables, which are all divided into multiple categories. Below all variables are discussed in detail, using information from the additional explanation to Article 6 of the Dutch Regulation risk equalization 2022 (Dutch Ministry of Health & Sport, 2021) and information presented by the Dutch National Health Care Institute (Zorginstuut, 2021). For each variable, the number of categories is mentioned. If this is stated as $X(+1)$, this means there are X categories plus one residual category, in which individuals are classified if none of the other X categories for this variable are applicable.

- Age interacted with gender (AG). This variable has 42 categories, divided into 21 male-based and 21 female-based groups. Within each gender, each category represents a five-year age group. Exceptions are made for new-borns and young adults. The age group 0-4 is split up in three categories: 0-years old (born in the year data is collected), 0-years old (born in the previous year) and 1-4 years old. The age group 15-24 is split up in a group of 15-17 and 18-24 years old. This is done as this distinguishes the adults, who are obliged to pay health care premiums in the Netherlands.

- Pharmacy-based cost groups (PCG). This variable is divided in 42(+1) subcategories, each of which represents the use of specific medicines corresponding to chronic diseases. The use of certain medicines above a pre-specified threshold in the preceding year classifies an insuree in such a category. An insuree can be classified in multiple categories. An extra category is created for people who do not use medicines corresponding to any of the 42 subgroups in this variable. This acts as a residual group and totals 43 subgroups.

- Diagnosis-based cost groups (DCG). This variable is based on the combination of diagnoses and treatments for chronic diseases in the preceding year. The variable consists of 26(+1) categories, which represent clusters of diseases with comparable cost patterns. An insuree can

both classify in a category multiple times as classify for multiple different categories. If none of the categories are applicable to an insuree, the insuree classifies for the residual group.

- Medical equipment based cost groups (MCG). This variable consists of 14(+1) categories, based on the use of medical equipment for chronic diseases. An insuree can be classified in multiple categories. There is a residual category for insurees who do not classify for any other category.

- Source of Income (SoI) interacted with age. This variable distinguishes insurees based on their source of income. Insurees from 70+ years of age are classified in a separate category. Insurees from 0 to 69 years are classified in either: fully incapacitated (with IVA benefit), incapacitated (no IVA benefit), social welfare assistance, student, self-employed, or higher education. Residual categories are present for people not classified in one these abovementioned categories. The categories are further split up in age subgroups (0-17 years, 18-34 years, 35-44 years, 45-54 years, 55-64 years, and 65-59 years), leading to 36 categories within this variable.

- Region. This variable consists of 10 categories, in which people are placed based on a two-step process. At first, socio-economic circumstances within a postal code are quantified using e.g. the percentages of low-income people. Next, based on this information postal codes are clustered in ten categories.

- Socio-economic status (SES) interacted with age. This variable is based on the total household income per address. The categories are: very low income, low income, middle income, and high income. Within these categories, three age groups (0-17 years, 18-69 years, and 70+ years) are distinguished which results in 12 categories.

- People per address (PPA) interacted with age. This variable classifies insurees based on the number of people living at an address. There is a separate category for children (0-17 years). Other people are divided into age groups (18-69 years, 70-79 years, and 80+ years) and classified in either: Wlz-institution (long-term stay), Wlz-institution (influx), 1-person households or a residual category. This variable contains 13 categories.

- Multiple year based high-cost groups (MHC). This variable is based on the somatic health care expenditure distribution of the last three years. It has 8(+1) categories. Category zero (the residual category) corresponds to insurees who were not present in the top 30% of the distribution in each of the past three years. Category one classifies people present in the top 30% of the distribution in one of the previous three years. Category two classifies people present in the top 10% of the distribution in the previous two years and category three to eight correspond to people who were among the top 15, 10, 7, 4, 1.5 and 0.5% of the health care expenditure distribution in each of the previous three years.

- Physiotherapy diagnosis-based cost groups (PDG). This variable indicates the use of physiotherapy for chronical diseases. It entails 4(+1) categories. Each category, except the residual category, represents a group of diagnoses.

- Multiple year based costs for nursing and caring (MNC). This variable contains 9(+1) categories. Category one to eight classify insurees present in the top 3.5%-0.25% of the distribution of health care expenditures on nursing and caring in each of the three preceding years. Category nine classifies children in the top 0.25% of the distribution in the previous year. Category zero represents the residual group.

- Historical somatic morbidity (HSM). This variable has 1(+1) category, based on historical health information. If three years ago an insuree classified for either a PCG, DCG, MCG, PDG, and/or MHC category (not being the residual category of each variable), the HSM variable is classified in this years' model. If not, insurees are classified in the residual category.

- Multiple year based pharmaceutical cost groups (MPC). This variable has 1(+1) category. Insurees classify for this variable if they were present in the top 25% of the distribution of pharmacy-related health care expenditure in at least one of the three preceding years. If not, they are classified in the residual category.

The total number of categories used in the 2022 model is 226. The number of categories used in the Dutch risk equalization model has increased over time, as is visible in table A1 in the Appendix. In the period 2019-2022, the number of categories for variables PCG, DCG, MCG, and SoI have changed. The variables HSM and MPC are added to the Dutch risk equalization model in 2022. Due to the large number of categories present in the Dutch risk equalization model and continuous improvements that are made, the model is widely regarded as one of the most extensive risk equalization models in the world.

The number of categories for the age interacted with gender (AG) variable normally is equal to 42. However, in this research subgroups of individuals are created based on the differences between observed and predicted annualized health expenditure in the previous three years. New-borns in 2017, 2018 and 2019 cannot be classified in one of these subgroups as they are not present in all years. They are therefore left out of the estimation process. In the resulting research sample, only new-borns from 2016 are present. Due to this choice, the age interacted with gender categories 1, 2, 22 and 23 cannot be classified as these correspond to male and female insurees of 0 years old in 2019.

2.3 Somatic health care expenditure in the Netherlands

In this section the somatic health care expenditure in the Netherlands for the period 2016-2019, which is used for the risk equalization in 2019 to 2022, is discussed. At first, descriptive statistics

for all years are displayed. Next, detailed information about the annualized somatic health care expenditure in 2019 is presented.

2.3.1 Actual somatic health care expenditure in 2016-2019

In table 1 on the next page, the descriptive statistics of the actual Dutch somatic health care expenditures over the period of 2016 until 2019 are displayed. This table is based on the Dutch population present in each of the four years. People for whom information in one of the years is missing (due to birth, death, or migration) are withdrawn from the data. This is done, as they can't be classified in one of the subgroups that are created in section 2.4. The unfiltered descriptive statistics, which represent the full population sample for each of the years, are presented in table A2 in the Appendix.

As is stated in the mentioned literature, health care data typically has a complex distribution. This distribution normally is characterized by a large peak of observations with 0 expenditure and a long right tail with large outliers. Such a large peak of observations with 0 expenditure is absent in this data set, as for each year the 1% percentile presents highly positive expenditures. This is the case, as this research focuses on somatic health care expenditure. In the Netherlands, different risk equalization systems are in use for somatic health care, mental health care and the own risk part of health insurance. It is typically the first variant, somatic health care, that does not display a large peak of observations with zero expenditure as this type of health care also incorporates the costs made at the general practitioner. These costs are made very commonly by people. In this research, this somatic health care expenditure is investigated.

However, in each of the four years visible in table 1 there is high positive skewness[3]. High positive skewness indicates a large right tail in the data, insurees with very high expenditure compared to the rest of the data set (Heij et al., 2004). This is illustrated by the large difference between the 99% percentile and the maximum health care expenditure in each of the years.

Next to this there is high kurtosis[4] which indicates many observations in the tails of the distribution (Heij et al., 2004). As kurtosis for each of the four years is very high, in each year the tails are thick. This is supported by information regarding the mean and the median in the data for each year. The mean health care expenditure in the period 2016-2019 increases from €1,988.94 to €2,455.13. However, the median ranges from €427.91 to €502.10 in this period. The mean is

[3]The third standardized sample moment for a distribution is called skewness. It is calculated as: $\frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^3}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^3}}$ (Heij et al., 2004), in which n is the total number of observations, y_i are the observations and $\bar{y}$ is the sample mean. Skewness measures the degree of symmetry of the data set. A symmetrical, normal distribution has skewness equal to 0. A data set with a large right tail indicates positive skewness.

[4]The fourth standardized sample moment for a distribution is called kurtosis. It is calculated as: $\frac{m_4}{s^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^4}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^4}}$ (Heij et al., 2004), using the same variables as used for the calculation of skewness. Kurtosis measures the magnitude of the tails of the distribution. Large kurtosis indicates thick tails. A normal distribution has a kurtosis around 3.

Table 1: Descriptive statistics of the actual Dutch somatic health care expenditures (in euros) in the period 2016 until 2019, for the merged data set.

	2016	2017	2018	2019
Individuals	16,327,282	16,327,282	16,327,282	16,327,282
Insured years	16,165,554.64	16,309,136.10	16,309,584.63	16,212,955.36
Mean	1,988.94	2,077.32	2,277.41	2,455.13
St. Dev.	6,679.02	6,905.77	7,572.88	8,074.31
Skewness	25.23	20.64	20.68	17.62
Kurtosis	2,712.42	1,344.67	1,926.62	1,029.30
Minimum	0.01	0.01	0.00	0.01
P1	63.59	65.81	68.25	71.53
Q25	156.51	159.20	169.81	182.19
Median	427.91	435.88	464.87	502.10
Q75	1,434.67	1,455.47	1,556.41	1,675.09
P99	25,793.76	27,591.26	30,830.25	33,556.27
Maximum	2,495,408.86	1,493,545.72	2,617,403.93	1,911,534.72

Table 1: The number of insured years is the sum of the weights corresponding to all individuals. P1 represents the first percentile of the actual health care expenditure distribution, Q25 and Q75 respectively the 25% and 75% quantiles and P99 represents the 99th percentile of the actual health care expenditure distribution.

shifted upwards heavily, due to a large amount of very large positive observations in the data set. In each of the years, this results in average actual health care expenditure which lies between the 75% and 99% percentiles.

2.3.2 Annualized somatic health care expenditure in 2019

In this research, the 2022 Dutch risk equalization system is of interest. This system is based on the 2019 health care expenditure data, scaled up to match the cost level in 2022. In table 1 a general overview of this data is visible. More detailed information is present in table 2, which is a selected part of table A3 visible in the Appendix.

In table 2 and A3 subsets of the population are created based on the variables used in the 2022 risk equalization model. Conditional on the created subsets, annualized health care expenditure data is displayed. Subsets that are created based on a cluster of categories within a variable include the applicable categories within brackets.

In table 2 it is visible that irrespective of the age, female insurees on average incur higher somatic health care expenditure in 2019 compared to male insurees. Next to this, the variables Multiple

Table 2: A detailed overview of the annualized somatic health care expenditures (in euros) for the 2022 risk equalization model.

Population subset	Individuals	Insured years	Frequency	Mean	St. Dev
Age interacted with Gender					
Female (<65 years)	6,436,063	6,417,502.45	39.58%	1,820.52	6,544.03
Female (>65 years)	1,827,657	1,792,789.68	11.06%	5,745.85	12,029.05
Male (<65 years)	5,944,821	5,918,553.90	36.51%	1,293.06	6,703.53
Male (>65 years)	2,118,741	2,084,109.32	12.85%	5,348.32	13,145.83
Multiple year high-cost groups					
No indication (0)	8,727,778	8,693,197.89	53.62%	794.92	3,585.19
In one year (1)	6,519,381	6,473,108.72	39.93%	2,948.46	7,838.89
In two years (2)	169,698	164,754.15	1.02%	10,608.96	19,766.61
In 3 years (3-8)	910,425	881,894.60	5.44%	14,786.24	22,921.41
Multi-year costs for nursing and caring					
No indication (0)	15,893,214	15,803,327.76	97.47%	2,212.93	7,345.36
Indication (1-9)	434,068	409,627.70	2.53%	17,799.55	25,105.94

Table 2: This table is a subset of the full table, visible in table A3. The number of insured years is the sum of the weights corresponding to all individuals. The frequency is calculated based on the relative number of insured years. The presented mean and standard deviation correspond to annualized somatic health care expenditure in 2019.

year based high-cost groups (MHC) and Multiple year based costs for nursing and caring (MNC) are displayed. Both variables classify insurees based on the level of health care expenditures in the previous three years. It is made visible in table 2 that the expenditures in years *t-3, t-2* and *t-1* are indicative for the expenditures in year t, as the higher categories within these variables display higher average annualized health care expenditure in 2019 as well.

2.4 Subgroups defined to evaluate model predictive performance

As mentioned in the literature, risk selection as performed by Dutch health insurers takes place on subgroup level. It is therefore of interest to evaluate the performance of the risk equalization models used in this research on subgroup level. In Table 3 the subgroups created for this evaluation are displayed.

The evaluated subgroups in this research are created based on the residual annualized health care expenditure in the period 2019-2021. Predicted annualized expenditure for each of these years is obtained via multiplication of the 2016-2018 variables in the data set (used in practice for the risk equalization of 2019, 2020 and 2021) with the corresponding, actual coefficients from the

Table 3: The subgroups used for evaluation, based on summed residual annualized expenditure for the period 2019-2021.

Subgroup	Aggr. Residual	UC	Individuals	Freq.	Mean res.	Median res.	St. Dev.
1	$AR > 0$	0	8,632,223	52.87%	3,578.78	2,260.20	5,323.29
2	$AR > 0$	1	3,580,598	21.93%	2,148.28	1,103.29	3,856.03
3	$AR > 0$	2	310,696	1.90%	1,480.46	508.58	3,184.39
4	$-5000 < AR < 0$	1	1,303,950	7.99%	-1,463.77	-1,008.81	1,353.75
5	$-5000 < AR < 0$	2	1,048,003	6.42%	-1,637.72	-1,186.53	1,400.10
6	$-5000 < AR < 0$	3	194,460	1.19%	-2,016.31	-1,663.36	1,363.38
7	$AR < -5000$	1	436,971	2.68%	-16,975.44	-9,844.95	35,085.97
8	$AR < -5000$	2	585,051	3.58%	-19,105.44	-10,773.16	39,290.76
9	$AR < -5000$	3	235,330	1.44%	-27,921.56	-13,881.81	57,663.10

Table 3: Aggregated residual (AR) measures the difference between the predicted and observed annualized somatic health care expenditure, aggregated over the period 2019-2021. UC measures the number of years that an insuree is undercompensated in the period 2019-2021. Frequency is measured as the relative number of individuals, present in each subgroup. For the aggregated residuals, the mean, median and standard deviation measured in euros are displayed.

risk equalization models of 2019-2021[5]. Residual annualized expenditure is equal to the difference between observed and predicted annualized expenditure. For each observation, the residuals for the period 2019-2021 are aggregated.

Based on these summed residuals, three subgroups are created: insurees with an aggregated overcompensation (AR$>$ 0), insurees with a small aggregated undercompensation ($-5000 <$ AR < 0) and insurees with a large aggregated undercompensation (AR$<$ -5000). These subgroups are further split up based on the number of years that an insuree is undercompensated within this period, ranging from 0 to 3 years.

The subgroups are created such that the main group of interest is clustered. Especially the insurees with high and persistent undercompensations over the period 2019-2021 are expected to be unwanted from the perspective of an insurer. It is this group of people, clustered in subgroups 8 and 9, that is most vulnerable to negative risk selection by an insurer.

Using the same argumentation, the individuals in persistently overcompensated subgroups 1 and 2 are most attractive for positive risk selection, from the perspective of an insurer.

[5]The coefficients from the risk equalization of 2019-2021 are obtained via WOR 930 (ESHPM, 2018), WOR 974 (ESHPM, 2019) and WOR 1002 (ESHPM, 2020).

3 Methodology

In this chapter, the methods and evaluation metrics used in this thesis are described in detail. At first the current Dutch linear risk equalization method, which is replicated in this thesis, is discussed. Thereafter the alternative models used in this thesis are explained. These models are combined into one combination of models, for which the procedure is explained. At last, the evaluation metrics used in this thesis are discussed.

3.1 Current risk equalization method: linear regression

The ex-ante compensation rewarded to health insurers for each individual in their portfolio is estimated via Ordinary Least Squares, with explanatory variables as discussed in section 2.2. The objective function is displayed in equation 1. This function is subject to constraints 2, 3 and 4.

$$\arg\min_{b} \sum_{i=1}^{N} w_i \left[y_i - \sum_{j=1}^{M} b_j x_{i,j} \right]^2 \tag{1}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} w_i \sum_{j=1}^{42} b_{AG(j)} x_{i,AG(j)} = \sum_{i=1}^{N} y_i \tag{2}$$

$$\sum_{i=1}^{N} w_i \sum_{j \in \kappa} b_{\kappa(j)} x_{i,\kappa(j)} = 0, \tag{3}$$

$$\kappa = \{PCG, DCG, MCG, Region, MHC, PDG, MNC, HSM, MPC\}$$

$$\sum_{i=1}^{N} w_i \sum_{Age,j \in \Psi} b_{\Psi(Age,j)} x_{i,\Psi(Age,j)} = 0, \tag{4}$$

$$\Psi = \{SoI, SES, PPA\}$$

The model is estimated without an intercept. w_i are the insured years, the fraction of the year an individual is insured in the Netherlands. y_i are the corresponding annualized somatic health care expenditures. b_j are the estimated coefficients to corresponding explanatory variables $x_j, j \in M$ with $M = 222$[6]. All explanatory variables are indicator variables, equal to either 0 or 1.

The constraints together ensure that in-sample total predicted annualized expenditure equals the in-sample total observed annualized expenditure. Constraint 2 constrains total in-sample annualized health care expenditure as predicted by variable 'Age interacted with gender' to equal total observed in-sample annualized health care expenditure. Constraint 3 is applicable to all other variables in the risk equalization model which are not interacted with age. For each of these variables, the sum of in-sample predicted annualized health care expenditure by those variable needs

[6]Note that in practice 226 categories are used. In this research, four age interacted with gender categories are ignored as discussed in section 2.2.

to equal 0. In-sample, these variables do not increase predicted annualized health care expenditure but redivide the predictions that are made by variable 'Age interacted with gender'. At last, constraint 4 is applicable to the variables 'Source of Income', 'Socio-economic status' and 'People per Address'. These variables are age-interacted. Constraint 4 ensures that within each age group of these variables, the sum of in-sample predicted annualized health care expenditure equals 0.

The application of OLS is highly interpretable on individual level, as the effect of an indication in a specific category is quantified by the corresponding coefficient. Each coefficient b_j is therefore interpreted as the ceteris-paribus change in predicted annualized somatic health care expenditure for an individual and thus the change in ex-ante compensation rewarded to an insurer, given that this indicator variable equals 1. However, there are downsides to this stringent model as well, as this method results in chronic under- and overcompensations of specific groups of insurees. Therefore, alternative methods to estimate the Dutch somatic health care expenditure are tested in this research. In the remainder of this chapter, these methods are explained in detail.

3.2 Regularized regression model

Regularized regression is a natural extension to linear regression to consider in this context. Given the large number of explanatory variables in the OLS model, risk of overfitting to the estimation data applies. Out-of-sample predictions made by an overfitted model typically display high variance and low bias, compared to underfitted models which typically display low variance and high bias. Regularized regression models respond to this bias-variance trade-off by decreasing model flexibility. This is done by means of a penalty term on the estimated coefficients (Hastie et al., 2017). Decreased model flexibility leads to reduced variance and increased bias for the predictions made by the model.

The most well-known types of regularized regression are Ridge regression, Elastic Net and Lasso regression. These models differ in terms of the penalty placed on the estimated coefficients. In Ridge regression, a penalty is placed on the squared values of the estimated coefficients. In Lasso regression, a penalty is placed on the absolute value of the estimated coefficients. The Elastic Net is a combination of both Ridge and Lasso regression, in which a variable determines what weight is placed on respectively the absolute and squared penalty term. The objective function of a regularized regression model is displayed in equation 5 (Hastie et al., 2017).

$$\arg\min_{b} \sum_{i=1}^{N} w_i \left[\left[y_i - b_0 - \sum_{j=1}^{M} b_j x_{i,j} \right]^2 + \lambda \left[\alpha \sum_{j=1}^{M} |b_j| + (1-\alpha) \sum_{j=1}^{M} b_j^2 \right] \right] \tag{5}$$

The parameters to tune within this function are α and λ. The value of α determines what type of regularized regression is performed. If $\alpha = 1$, the penalty on the absolute value of the coefficients is activated and thus Lasso regression applies. If $\alpha = 0$, the penalty is placed on the squared coefficients leading to Ridge regression. Each value of α between 0 and 1 applies to Elastic Net, a

combination of both Ridge and Lasso regression. The magnitude of the penalty term is determined by λ. A higher λ indicates a larger penalty term. Note that, when $\lambda = 0$ equation 5 reduces to equation 1, a simple linear regression. In this research, the optimal values of α and λ are determined based on a grid search, as discussed in section 4.1.

3.3 Conditional Density Approximation using Ordered Logit

As opposed to the widely used two-part models to estimate the probability to have positive health care expenditure, the extended two-part model which estimates the probability of insurees to fall within a pre-defined cost interval (Gilleskie & Mroz, 2004) is not yet applied regularly, Jones et al. (2015) note. Given the promising results in literature (Gilleskie & Mroz, 2004; Jones et al., 2015) combined with the high interpretability of the model, this technique provides an attractive alternative to the current OLS model.

The first step of this conditional density approximation is to create an ordinal categorical variable, as this is the only type of dependent variable an Ordered Logit model can be estimated on. Therefore, intervals of insurees are created based on annualized health care expenditure level in 2019. Individuals in the merged data set are ordered ascendingly based on the invididual level of annualized expenditure in 2019. Thereafter, intervals of insurees are clustered together. Each interval is responsible for a percentage of the total somatic health care expenditure in the Netherlands in 2019. For example, given 10 intervals the total annualized expenditure within each interval corresponds to 10% of the total somatic health care expenditure in the full sample.

The number of intervals K in the model estimation is of great influence, as different numbers of intervals cluster different individuals together. This changes the mean and median annualized expenditure within these intervals, used in the next step of this model. Note that for K created intervals, equation 6 with N_k the number of individuals in each interval holds. Therefore, especially the higher intervals with less individuals in it, are heavily influenced by the chosen number of intervals K. A low number of intervals creates highly heterogeneous clusters of insurees, a high number of intervals results in very small group sizes N_k within intervals.

$$N_1 \geq N_2 \geq ... \geq N_K, k = 1, ..., K \tag{6}$$

Gilleskie and Mroz (2004) indicate that the use of 10 to 20 intervals to partition the data in presents good results. Jones et al. (2015) use 15 intervals, each of which contain an equal number of individuals, to partition the data in. In this research, different numbers of intervals are tested. Using 5-fold cross validation, the optimal number of intervals K is chosen based on a grid search among 10, 12, 15, 17 and 20 intervals. Given the interval indication as the ordinal dependent variable, for each individual i the Ordered Logit regression model estimates the probabilities $p_i(k)(X_i)$ to be classified in each interval.

The resulting probabilities $p_i(k)(X_i)$ for $k = 1, ..., K$ are used to predict the annualized somatic health care expenditure for individuals by means of equation 7 (Jones et al., 2015) and 8.

$$E\left[y_i|X_i\right] = \sum_{i=1}^{K} p_i(k)(X_i)\bar{y_i} \tag{7}$$

$$E\left[y_i|X_i\right] = \sum_{i=1}^{K} p_i(k)(X_i)\tilde{y_i} \tag{8}$$

Predictions are made both on mean and median annualized health care expenditure within intervals. This is done as the mean expenditure is expected to be influenced heavily by high-cost individuals within each interval. This is not the case for median expenditure. By incorporating both metrics in the model estimation, the effect of these high-cost individuals on the model outcome can be assessed.

3.4 Regression Tree models

A class of models found to be very promising regarding the literature are regression tree models (Rose, 2016; Stam et al., 2020; Vimont et al., 2022). In this section, the conceptual idea of regression trees is explained. Next, two extensions of regression trees are introduced for this research: a Random Forest, and a Gradient Boosted model. The models are explained using formulas presented by Hastie et al. (2017).

Regression tree models rely on recursive binary splits of the data into disjoint subsets. The binary splits are based on splitting variables and points as illustrated in equation 9 (Hastie et al., 2017). In this equation, x_j is the variable chosen to partition on and s is the partition point. All observations that respectively meet and do not meet the condition are clustered together into the subsets S_1 and S_2.

$$S_1(j, s) = \{x|x_j \leq s\} \quad S_2(j, s) = \{x|x_j > s\} \tag{9}$$

This choice of variable x_j and s is based on the minimization of the sum of loss functions $L\left[y_i, \hat{f}(x_i)\right]$ over the subsets created by the data partition, with $\hat{f}(x_i)$ the predicted value of y_i based on the resulting regression tree. Given a squared error loss function, this results in equation 10 (Hastie et al., 2017), in which the aggregated sum of squared residuals over the two created subsets S_1 and S_2 is minimized. To find the optimal binary split the regression tree algorithm evaluates all possible combinations of x_j and s. For the data split as presented in equation 10, the optimal value assigned to each observation within the two subsets are equal to γ_1 and γ_2. Finding these values for each data split is an optimization problem as well.

For each possible data partition into subsets S_1 and S_2, taking the first derivative of equation 10 with respect to γ_1 and γ_2 and setting these equal to 0 yields the optimal values for γ_1 and γ_2,

displayed in equation 11 (Hastie et al., 2017). The optimal values are equal to the conditional mean within the respective created subsets of the data.

$$\arg\min_{j,s}\left[\arg\min_{\gamma_1}\sum_{x_i\in S_1(j,s)}\frac{1}{2}(y_i-\gamma_1)^2+\arg\min_{\gamma_2}\sum_{x_i\in S_2(j,s)}\frac{1}{2}(y_i-\gamma_2)^2\right] \tag{10}$$

$$\gamma_a=\bar{y}_{S_a}=E\left[y_i|x_i\in S_a(j,s)\right], a=1,2 \tag{11}$$

The data partitioning procedure as displayed in equation 10 and 11 can be repeated multiple times, to further divide the created disjoint subsets of observations into smaller and more specific disjoint subsets. The data partitioning is stopped when an a-priori defined stopping condition is met.

As indicated by Hastie et al. (2017) the outcome of the fully fitted regression tree is equal to equation 12, in which Θ captures the parameters $\{S_a,\gamma_a\}_{a=1}^{A}$, each corresponding to one of the A created disjoint subsets of the data, called final nodes.

terminal nodes of the regression tree in equation 12 (Hastie et al., 2017). The values for $\gamma_a, a = 1, ..., A$ are obtained via equation 11.

$$T(x;\Theta)=\sum_{a=1}^{A}\gamma_a\mathbf{1}_{\{x_i\epsilon S_a\}} \tag{12}$$

Regression trees are highly interpretable. The model explained in this section can easily be visualised by means of a 2D-graphic which illustrates how each terminal node is constructed. Next to high interpretability, a single regression tree generally has a low bias (Hastie et al., 2017). This originates from the fact that it does not make assumptions on the distribution of the data. A regression tree can therefore fit the training data very well. However, this comes at the cost of high variance. In the procedure of fitting the model, the binary splits of a regression tree are highly dependent on the training data, a change in training data could therefore have major impact on the outcomes of the regression tree.

Extensions on this technique include the use of a large number of separately build trees, merged into one prediction model to reduce this variance. Two of these extensions, a Random Forest, and a Gradient Boosted model, are used in this research.

3.4.1 Random Forest

A Random Forest, as first introduced by Breiman et al. (2001), is a bootstrap aggregated variant of regression trees. This method grows multiple regression trees and averages out the predictions from all trees into one final prediction. The final prediction of a Random Forest is displayed in equation 13 (Hastie et al., 2017), with $T(x;\Theta_k)$ the predictions from regression tree k as indicated in equation 12.

$$\hat{f}_{RF}(x) = \frac{1}{K}\sum_{k=1}^{K} T(x;\Theta_k) \tag{13}$$

Each regression tree in the Random Forest model is estimated using bootstrapping (sampling with replacement) to draw a random sample of observations from the full data set used in the estimation process. Furthermore, only a subset of all variables is presented to each separately grown tree in the Random Forest. These restrictions are imposed on the regression trees to reduce the variance of the predictions from equation 13. This variance is equal to (Hastie et al., 2017):

$$Var\left[\frac{1}{K}\sum_{k=1}^{K} T(x;\Theta_k)\right] = \rho\Sigma + \frac{1-\rho}{K}\Sigma \tag{14}$$

With ρ the correlation between separately grown trees and Σ the covariance matrix. As the number of trees K within the Random Forest grows large, the second term of equation 14 converges to 0. This leaves only the first term, dependent on the correlation between the trees. The restrictions regarding bootstrapping and the consideration of only a subset of variables for each data split are used to lower this correlation between trees and thus to lower the variance of the Random Forest predictions.

3.4.2 Gradient Boosted Regression Tree

Gradient Boosting relies on a recursively build set of regression trees. In each iteration of the algorithm, a tree is added to the current version of the model. Tree iteration m is hereby generated based on the optimization of $\hat{\Theta}_m$ for the loss function in equation 15 (Hastie et al., 2017).

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} L\left[y_i, (f_{m-1}(x_i) + \eta T(x_i;\Theta_m))\right] \tag{15}$$

Here, $f_{m-1}(x_i)$ equals the latest version of the model evaluated at x_i, the result of iteration $m-1$ of the algorithm. Regression tree $T(x_i;\hat{\Theta}_m)$ is added to this in step m, which results in a new version of the model. The magnitude of the effect of tree m on model *m-1* is determined by the learning rate η. This learning rate is the model adaptation to each new iteration. An increased learning reate results in larger updates to the existing model by each new iteration.

The fastest decrease in the value of the loss function in equation 15 can be obtained by taking the negative gradient of the loss function with respect to the prediction model (Hastie et al., 2017). Using a squared error loss function, Hastie et al. (2017) indicate that this negative gradient evaluated through all predictions of model f_{m-1} is equal to the residuals of the predictions made by model f_{m-1}, as displayed in equation 16 (Hastie et al., 2017).

$$-\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=f_{m-1}(x_i)} = -\left[\frac{\partial \frac{1}{2}(y_i, f(x_i)^2}{\partial f(x_i))}\right]_{f(x_i)=f_{m-1}(x_i)} = y_i - f_{m-1}(x_i) = e_{i,m-1} \quad (16)$$

Therefore, to optimize equation 15 the regression tree corresponding to iteration m of the Gradient Boosted algorithm should be fitted to the residuals of the preceding prediction model f_{m-1}. This translates into equation 17 (Hastie et al., 2017), in which the sum of squared errors between the residuals of model $m-1$ and tree $T(x_i; \Theta_m)$ is minimized:

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} (e_{i,m-1} - T(x_i; \Theta_m))^2 \quad (17)$$

The regression tree $T(x_i; \hat{\Theta}_m)$ resulting from equation 17 is then added to model f_{m-1}, using the learning rate η. This creates the updated model f_m. This process is iterated M times, resulting in the final model f_M.

3.5 Combined model: stacked generalization

All four predictive models discussed in sections 3.2 to 3.4 are evaluated against the current linear risk equalization model. Next to this a linear combination of the four models is generated. This is called a stacked model (Wolpert, 1992). This model builds upon the idea of a combination of forecasts as introduced by Bates and Granger (1969). Predictions made by the stacked model are in essence the outcome of a two-stage process: predictions are made by the underlying models and these predictions are weighted by the fitted stacked model. Given the use of estimation models in both the first and second stage of the stacked model, it is of high importance to take caution on the use of data and avoid overfitting on the data used in the estimation process. Therefore, in this section both the use of data and the estimation process of the stacked model are discussed.

3.5.1 Optimal model specifications of the underlying models

As outlined in the data section of this research, the data set with merged information over 4 years contains 16,327,282 individuals. These insurees are present in the full period 2016-2019. This data is randomly split in two parts using an 80/20 division which results in the learning set (13,060,528 individuals) and the test set (3,266,754 individuals). The learning set is used to train and validate the models in this research. The test set is used to evaluate the predictive performance of the models on out-of-sample data.

Using the learning set to train the models in this research, the first step is to decide on the optimal model specifications. Hyperparameters that require tuning are displayed in table 4.

Table 4: Hyperparameters tuned to determine optimal model specifications.

Model	Hyperparameter	Explanation
Regularized regression	α	Determines the type of regularized regression.
	λ	Determines the magnitude of the penalty term.
Ordered Logit	K	Number of cost intervals.
Random Forest	Trees	Number of trees.
	Node size	Minimum node size of each final node in a tree.
	Mtry	Number of variables considered in each data split.
Gradient Boosted model	Iterations	Number of iterations.
	Depth	Number of data splits in each tree.
	η	Learning rate.

Table 4: Hyperparameters of the models used in this research that require tuning.

The hyperparameters are tuned based on 5-fold cross validated predictive performance for a subset of 2,000,000 observations from the learning set[7]. This subset of 2,000,000 observations is validated to be representative for the full learning set in terms of the distribution of annualized expenditure, as indicated by table A4 in the Appendix. The metrics used to measure predictive performance are introduced in section 3.6.

3.5.2 Combined model using stacked generalization

Given the cross-validated optimal model specifications of each method, the models are trained in two ways. This training process is displayed in figure 1. Using an 80/20 division, the learning set is randomly split again into the training set (10,450,146 observations) and the validation set (2,610,382 observations). The models used as input for the stacked model are estimated on the training set. Predictions based on these estimated models are made for the validation set. These hold-out validated predictions for annualized somatic health care expenditure are denoted as $\hat{F} = \{\hat{f}_{RR}, \hat{f}_{OL}, \hat{f}_{RF}, \hat{f}_{GB}\}$, where RR, OL, RF and GB respectively correspond to Regularized regression, Ordered Logit, Random Forest and Gradient Boosted model.

The combined model is estimated via OLS of the annualized somatic health care expenditure on the set of predictions $\hat{F}$. The estimated coefficients are weights, they weigh the predictive outcomes of the single predictive models into one final prediction. This combined model is called a stacked model, as introduced by Wolpert (1992).

The resulting stacked model is thus estimated using both the training and the validation set, i.e.

[7] A subset of the full learning set is used for hyperparameter tuning, as this substantially reduces runtime of the models. This research uses confidential data and therefore is performed on an isolated server, which has limited computational capacity. Therefore, this choice based on computational feasibility is made.

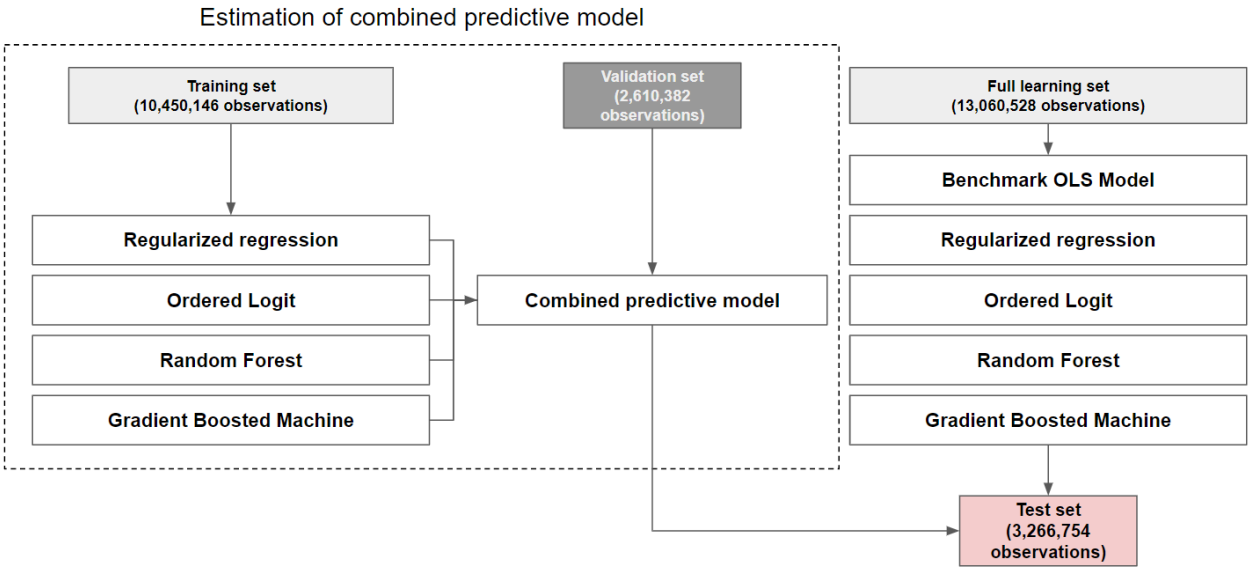

Figure 1: Overview of the final estimation procedure of the models used in this research.

Figure 1: The estimation process of the final models in this research. Prior to this, optimal model specifications for the Regularized regression, Ordered Logit regression, Random Forest model and Gradient Boosted model are determined using 5-fold cross validated predictive results on a subset of 2,000,000 observations of the full learning data set.

the full learning set. The OLS model as discussed in section 3.1 is trained on the full learning set as well. A fair comparison of the stacked model and OLS model to the other models in this research, therefore requires the other models to be trained once on the full learning set as well. This results in 6 models considered in this research, trained on the same set of data. These models are evaluated using the test data. This data is completely new to each model, not used earlier in the estimation process.

3.6 Evaluation of predictive performance

An important first note in the evaluation of the models, is that expenditure as predicted by all models is rescaled such that total predicted annualized expenditure by each model equals total observed annualized expenditure for the data set for which predictions are made. Risk equalization in essence is a redistribution of the risks over health insurers and thus each model should be assumed to redivide the same amount of money over health insurers. The use of rescaled predictions such that total annualized predicted expenditure is equal among all models is good practice in risk equalization literature.

As indicated in the introduction, one of the purposes of risk equalization is to reduce incentives for risk selection. Risk selection takes place on subgroup level and thus the evaluation of risk

equalization models should also be based on non-random subgroups (Van Veen et al., 2015), as created in section 2.4 of this research.

The use of these subgroups however makes it hard to evaluate overall performance of the models (Van Veen et al., 2015). For example, in the parameter tuning process one model specification can perform good on a certain subgroup, but another specification can present favourable results for another subgroup. Therefore, the models are evaluated both on individual level and on non-random subgroup level. Optimal model specifications are determined based on individual-level metrics. In this section, the metrics used to evaluate the predictive performance of the models are discussed.

3.6.1 Individual level evaluation metrics: R^2, CPM and MSE

R^2 is the most popular metric (Van Veen et al., 2015). It uses weighted squared differences between predicted and observed annualized expenditure relative to the weighted squared differences between observed and mean observed annualized expenditure, as outlined in equation 18 (Heij et al., 2004). The weights w_i correspond to the fraction of the year, an insuree is registered. The R^2 indicates the predictive performance of an algorithm in terms of total variance explained by the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{N} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} w_i (y_i - \bar{y})^2} \tag{18}$$

The `CPM` (Cummings Prediction Measure) as introduced by Cumming et al. (2002) is similar to R^2, but uses the weighted absolute differences instead of weighted squared differences, as outlined in equation 19 (Cumming et al., 2002). This is informative to compare with R^2 values, as the R^2 is based on squared differences and thus heavily influenced by large differences between predicted and observed annualized expenditure. This magnified effect of large differences is not present using absolute differences.

$$CPM = 1 - \frac{\sum_{i=1}^{N} w_i |y_i - \hat{y}_i|}{\sum_{i=1}^{N} w_i |y_i - \bar{y}|} \tag{19}$$

Both the R^2 and `CPM` display outcomes between 0 and 1, hence they are called standardized evaluation metrics (Van Veen et al., 2015). For both metrics, a value closer to one indicates higher predictive performance.

Next to these standardized individual level metrics, the non-standardized weighted mean squared error (MSE) is calculated by equation 20. As this metric presents the weighted average squared residual of the predictions of a model, a lower MSE value indicates higher predictive performance.

$$MSE = \frac{\sum_{i=1}^{N} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} w_i} \tag{20}$$

3.6.2 Subgroup level evaluation metric: MPE

To evaluate subgroup level predictive performance, the mean prediction error (MPE) within each subgroup as defined in section 2.4 is calculated in equation 21.

$$MPE = \frac{\sum_{i=1}^{N} w_i(\hat{y}_i - y_i)}{\sum_{i=1}^{N} w_i} \tag{21}$$

Given that this metric uses untransformed differences between observed and predicted annualized health care expenditure, positive and negative differences within a subgroup cancel out. Therefore, this metric does not present a clear insight in the accuracy of a model. It is however very informative, as for a health insurer the MPE of a subgroup equals the expected gain or loss, obtained by incorporating an individual from a subgroup in their portfolio.

Note that the ordering of terms in the numerator of equation 21 is changed compared to the numerators of the equations for R^2, CPM and MSE. As the other metrics use absolute or squared differences, the ordering of the terms does not matter. In the calculation of MPE the values are untransformed and thus, the ordering does matter. Changing the ordering of the terms leads to a neater interpretation of the MPE value for subgroups. Using equation 21, a negative MPE value corresponds to lower predicted annualized expenditure compared to observed annualized expenditure, and thus an expected undercompensation. A positive MPE value indicates higher predicted annualized expenditure and thus an expected overcompensation.

4 Results

In this chapter, the results of the estimated models are presented. At first, the optimal model specifications for the Regularized regression, Ordered Logit regression, Random Forest model and Gradient Boosted model are determined based on the cross validated predictive results for the subset of 2,000,000 observations as discussed in section 3.5.1. Next, the definitive models are estimated and stacked into one final predictive model. The predictive results of this stacked model and the single predictive models trained on the full learning set are used as an evaluation benchmark against the current OLS risk equalization model in the Netherlands.

4.1 Regularized regression model

The regularized regression model requires optimization of two hyperparameters. A grid search is performed for parameter α on the interval $\{0.0, 1.0\}$ with steps of 0.1 (11 model specifications). This results in a spectrum of one Lasso regression, nine Elastic Net regressions and one Ridge regression. For each of these models, λ values ranging from 0 to 300 (with step size 2) are evaluated. The optimal λ is chosen based on the minimal cross-validated mean squared error within the set of models for a given α. The results of this grid search are displayed in table 5.

Table 5: Cross-validated predictive results for model specifications of the regularized regression.

α	λ	MSE	R^2	CPM
1.0 - 0.3	0	49,281,428	0.366	0.352
0.2	2	49,279,731	0.366	0.353
0.1	10	49,276,433	0.366	0.353
0	254	49,232,163	0.367	0.352

Table 5: Optimal model specifications of the regularized regression for each level of α and the corresponding cross-validated predictive performance of these models. The chosen model specification is indicated in grey.

Results for $\alpha = 0.3$ to $\alpha = 1.0$ are clustered together, as for each of these model specifications an optimal λ of 0 is obtained. Referring to equation 5 in section 3.2, regularized regression with $\lambda = 0$ reduces to Ordinary Least Squares. Therefore, the optimal model specifications for $\alpha = 0.3$ to 1.0 are identical OLS models. The observed cross-validated predictive results for these models differ by very small amounts due to randomization in the cross-validation process. However, as each of these models essentially is the same the results are grouped in one row. The best cross-validated predictive outcome among these 8 OLS models is displayed in table 5.

For the Elastic Nets with $\alpha = 0.2$ and $\alpha = 0.1$ a non-zero optimal value of λ is obtained, as well as for the Ridge regression ($\alpha = 0$). Therefore, in these models the penalty term is activated

leading to different model specifications and outcomes. The regularized regression with highest predictive performance in terms of cross-validated MSE and R^2 is the Ridge regression. Despite slightly inferior predictive results in terms of `CPM` compared to the Elastic Net models, the Ridge regression with $\lambda = 254$ is therefore chosen as the model specification and thus used further in this research.

4.2 Ordered Logit model

As discussed in section 3.3, the predictive outcome of the Ordered Logit model highly depends on the number of cost intervals K the data is partitioned in. In this section, a grid search among $K = \{10, 12, 15, 17, 20\}$ is conducted. For each of these number of intervals, predictions are made both using mean and median annualized health care expenditure within these intervals. In table 6 the cross-validated individual level predictive performance for each evaluated model specification is displayed.

Table 6: Cross-validated predictive results for model specifications of the Ordered Logit regression.

Intervals	Metric	MSE	R^2	CPM
10	Mean	56,900,958	0.265	0.355
10	Median	54,388,214	0.298	0.386
12	Mean	58,814,791	0.241	0.355
12	Median	55,528,535	0.283	0.381
15	Mean	60,756,222	0.216	0.353
15	Median	57,121,373	0.262	0.372
17	Mean	68,596,383	0.114	0.332
17	Median	62,533,701	0.193	0.351
20	Mean	79,094,730	0.020	0.279
20	Median	69,978,620	0.096	0.299

Table 6: Cross-validated predictive results for the grid of Ordered Logit models. The chosen model specification is indicated in grey.

The use of median annualized expenditure compared to mean annualized expenditure improves the predictive performance for each of the evaluated interval levels. Next to this, for each of the evaluated models the `CPM` value is considerably higher compared to R^2. This indicates a set of large residuals in the predictions made by the models, as these are weighted more heavily in the computation of R^2 and therefore penalize predictive performance in terms of this measure. Next to this, an increase of the number of intervals deteriorates the predictive performance of the models. Based on the results displayed in table 6, the Ordered Logit model with $K = 10$ intervals and the use of median expenditure is chosen to work with in this research.

4.3 Random Forest

For a Random Forest, hyperparameters that require tuning are the number of trees in the algorithm, the minimum number of observations required in every final node of each tree and the number of variables (Mtry) considered for each data splitting rule within a regression tree. In table 7 the different model specifications are displayed. Each hyperparameter of the Random Forest algorithm is tuned in a different round of iterations.

In the first round, Random Forests with 100 trees and 14 variables[8] to consider for each data split are fitted. For these models, minimum node size is varied from 5 to 100. In this round, the models with a minimum node size of 15 and 50 display the highest predictive performance. These models are fitted again in the second round with the number of trees increased to 500. In this round, the model with a minimum of 50 observations per final node performs best on two of the three evaluated metrics. In the last round this model is fitted again with different numbers of variables considered for each data split in a regression tree. This results in highest predictive performance obtained by a Random Forest with 500 trees, a minimum of 50 observations in each final node and 20 variables considered per data split. This model specification is used further in this research.

Table 7: Cross-validated predictive results for model specifications of the Random Forest model.

Round	Trees	Node size	Mtry	MSE	R^2	CPM
1	100	5	14	50,134,327	0.355	0.353
	100	15	14	49,976,245	0.357	0.352
	100	50	14	49,969,284	0.357	0.351
	100	100	14	50,075,799	0.356	0.350
2	500	15	14	49,825,642	0.359	0.353
	500	50	14	49,793,347	0.360	0.352
3	500	50	10	50,084,197	0.356	0.349
	500	50	20	49,639,143	0.362	0.353
	500	50	30	49,696,227	0.361	0.353

Table 7: Cross-validated predictive results for each of the evaluated model specifications of the Random Forest model. Node size displays the minimum number of observations required in each final node of a regression tree, Mtry equal the number of variables considered for each data split in a regression tree. The chosen model specification is indicated in grey.

[8]In the first 2 rounds of hyperparameter tuning of the Random Forest model, the number of variables considered in each data split is set to 14. This is equal to the square root of the total number of explanatory variables (M=222), a rule of thumb normally applicable to classification trees. For regression trees, the rule of thumb is to consider M/3 variables per data split. Deviation from this rule of thumb is motivated by limited computational capacity. In the third round of hyperparameter tuning the number of variables is varied. Here is shown that the initial chosen value of 14 is not far apart from the optimal value found, equal to 20.

4.4 Gradient Boosted model

For the Gradient Boosted model the number of iterations, the depth of each fitted tree in an iteration and the learning rate need tuning. The depth is equal to the number of data splits performed in each regression tree. The learning rate is the model adaptation to each new iteration, denoted by η. The parameter tuning process of the Gradient Boosted model is conducted in four rounds, as displayed in table 8.

In the first round, a model with 100 iterations and $\eta = 0.2$ is fitted. These are arbitrary choices, as these parameters are tuned at a later stage. The depth of each tree in the algorithm is differed from 1 to 5. Highest predictive performance is obtained by a depth of 3, followed by a depth of 1. In the next two rounds of the tuning process, these two best performing models from round 1 are repeated with an increased number of iterations. In terms of both MSE and R^2, a tree depth of 1 displays the best results for the models with 250 and 500 iterations. This model specification is fitted again, varying the learning rate from 0.1 to 0.4. Presented with the cross-validated predictive results, the Gradient Boosted model with 500 iterations, a depth of 1 and learning rate 0.3 performs best on both MSE and R^2. Despite small outperformance by other model specifications in terms of CPM, this model is used further in this research.

Table 8: Cross-validated predictive results for model specifications of the Gradient Boosted model.

Round	Iterations	Depth	η	MSE	R^2	CPM
1	100	1	0.2	51,535,502	0.337	0.315
	100	3	0.2	50,285,079	0.353	0.343
	100	5	0.2	51,847,113	0.333	0.350
2	250	1	0.2	49,968,953	0.357	0.337
	250	3	0.2	50,762,900	0.347	0.351
3	500	1	0.2	49,496,662	0.363	0.346
	500	3	0.2	51,479,551	0.338	0.354
4	500	1	0.1	50,017,480	0.357	0.335
	500	1	0.3	49,355,249	0.365	0.350
	500	1	0.4	49,397,414	0.361	0.353

Table 8: Cross-validated predictive results for each of the evaluated model specifications of the Gradient Boosted model. Depth equals the number of data splits performed in each regression tree, η is the learning rate, the model adaptation to each new iteration. The chosen model specification is indicated in grey.

4.5 Stacked predictive model

For the stacked predictive model, the selected model specifications from sections 4.1 to 4.4 are fitted on the training set. These fitted base models are used to predict annualized somatic health care expenditure for individuals in the validation set. These validation set predictions for the base models are used as explanatory variables for the final stacked model, in which the annualized somatic health care expenditure is regressed on these predictions by means of a multiple linear regression. The model estimation output is displayed in table 9.

Table 9: Model estimation results of the stacked multiple linear regression model.

Model	Coefficient	Standard deviation	t-value	p-value
Ridge regression	0.042	0.005	8.38	< 0.01
Ordered Logit	-0.024	0.002	-11.36	< 0.01
Random Forest	0.467	0.005	103.27	< 0.01
Gradient Boosted	0.532	0.005	113.84	< 0.01

Table 9: Model estimation results of the stacked model, using validation set predictions of the other models as explanatory variables.

In table 9 it is visible that each of the estimated coefficients are significantly different from 0 at 1% significance level. Next to this, both the predictions from the Ordered Logit model as from the Ridge regression model receive a very small coefficient compared to the Random Forest and Gradient Boosted model predictions. Therefore, the predictions made by the stacked model are mostly influenced by the Random Forest and Gradient Boosted model. Predictions from the Ordered Logit model are even weighted negatively. Note that predictions from the Ordered Logit model are always positive, as they are the outcome of predicted probabilities to fall in certain intervals multiplied with median expenditure within these intervals. Both these values are non-negative. Therefore, annualized somatic health care expenditure predictions from the stacked model are decreased by the predictions from the Ordered Logit regression.

4.6 Evaluation of predictive performance

In this section, the predictive performance of the current OLS risk equalization method and the alternative models is presented. All predictive models evaluated in this section are estimated using the full learning set, whereby the stacked model uses the training set to estimate base models and the validation set predictions from these base models to estimate the coefficients. This training and validation set together form the learning set.

At first, individual level performance of the models is displayed. Next, the subgroup level performance as well as the variance of the predictions is presented. Note that the displayed results

correspond to rescaled predicted expenditure by each model, to match total observed annualized health care expenditure level. The unscaled final results of this research are presented in table A5 and A6 in the Appendix.

4.6.1 Individual level predictive performance

The individual level predictive results are displayed in table 10. Note that both the Ridge regression as the Ordered Logit regression fail to outperform OLS regression on any of the evaluated metrics. The Random Forest, Gradient Boosted model and stacked model do outperform the OLS regression on all metrics, albeit with small amounts. The stacked model displays the highest predictive performance of all models, which indicates that this model fits the data best.

Table 10: Individual level predictive performance of the evaluated models for the test set.

Model		MSE	R^2	CPM
Base models	Ridge regression	47,609,155	0.371	0.374
	Ordered Logit	54,800,557	0.276	0.366
	Random Forest	47,183,396	0.377	0.381
	Gradient Boosted model	46,977,305	0.379	0.380
Stacked model		46,651,273	0.384	0.382
OLS regression		47,319,674	0.375	0.377

Table 10: Predictive results on the test set, for the final models fitted using the full learning set. Model specifications for the base models are chosen as outlined in section 4.1 to 4.4.

4.6.2 Subgroup level predictive performance

In table 11 the predictive results on subgroup level are displayed, next to the observed mean annualized expenditure within each subgroup. The MPE values for each regression model display the mean deviation of predicted annualized expenditure from observed annualized expenditure. Positive MPE values indicate an average overcompensation compared to the observed annualized expenditure level, negative MPE values indicate an undercompensation.

Note that subgroups 1 to 9 are created based on the aggregated results on risk equalization[9] for the risk equalization models of 2019 to 2021, as discussed in section 2.4. Subgroups 1 to 3 display a mean aggregated positive result on risk equalization over the years 2019 to 2021, which indicates a historic overcompensation for individuals in these subgroups. Subgroups 4 to 9 display negative historic results, which indicate a historic undercompensation. For individuals in subgroups 7 to 9 this undercompensation was severe (>€5000).

[9]The result on risk equalization is a term used in risk equalization literature. It represents the residual between observed and predicted annualized health care expenditure for each individual.

Table 11 displays that individuals present in these severely undercompensated subgroups 7 to 9 on average have considerably higher annualized somatic health care expenditure in 2019 compared to subgroups 1 to 6. Highest mean annualized somatic health care expenditure in 2019 is found for subgroup 9, the subgroup which consists of individuals that are undercompensated in each of the past three years and whose undercompensation adds up to more than €5000.

Next to this, the indicative value of historic information can be assessed. Except for the mean prediction by the Ordered Logit for individuals in subgroup 8, all models predict an average undercompensation for individuals in subgroups 7 to 9 in this year as well. Furthermore, historically overcompensated individuals in subgroups 1 and 2 are overcompensated by all predictive models in this year as well, except for the mean prediction by the Ordered Logit model for subgroup 1. The indicative value of the historic results of the risk equalization model, on which the subgroups are created, therefore is present in this data.

Table 11: Subgroup level predictive performance in terms of MPE (in euros) of the evaluated models for the test set.

Group	Observed	Ridge	Ord. Logit	RF	GBM	Stacked	OLS
1	1,523.93	120.28***	-5.73***	101.77	103.10	105.84	105.91
2	2,406.67	90.88*	160.92***	89.43*	83.53***	84.76**	107.27
3	3,430.59	-35.01	165.37***	-38.94	-45.36	-46.86	2.62
4	2,335.69	-32.35**	-45.43***	-12.47	-15.52	-14.22	5.30
5	2,469.93	-258.32***	-196.86	-212.60	-228.33**	-223.28*	-193.98
6	1,981.34	-470.56*	-477.52**	-407.56	-411.73	-411.04	-416.90
7	7,626.01	-154.58***	-331.51	-173.90**	-149.02***	-156.39***	-437.93
8	9,751.06	-680.42	165.00***	-630.04*	-622.37**	-647.02*	-749.24
9	14,542.22	-2126.49***	-745.97***	-1868.82	-1791.34	-1866.94	-1744.37

Table 11: Subgroups 1-9 are defined as in Table 3 in section 2.4. The column with 'Observed' values presents observed mean annualized somatic health care expenditure within each subgroup. In the column titles, Random Forest and Gradient Boosted model are abbreviated with respectively RF and GBM. MPE values indicated with (*), (**) or (***) differ significantly from the MPE values obtained by the OLS model on respectively 10%, 5% or 1% significance level as tested by means of a two-sample t-test.

Note that in table 11, for each subgroup MPE values close to 0 are desired. Values close to 0 indicate, on average, a small residual between observed and predicted annualized expenditure within a subgroup. In that case, there is no clear incentive for positive or negative risk selection on this subgroup from the perspective of an insurer.

For subgroups 3, 4, 5, and 6 there is no significant outperformance of the OLS model. The

MPE values of the alternative models for these subgroups that are significantly different from those obtained by the OLS model display significantly worse predictions for the alternative models compared to the OLS model, as the MPE values are further away from 0.

For subgroups 1, 8 and 9, the Ordered Logit model displays significantly better results compared to the OLS model. It predicts a small undercompensation for subgroup 1 instead of a high overcompensation as is done by all other models, including the OLS model. For subgroup 9 it predicts a significantly lower overcompensation compared to the OLS model. However, the Ordered Logit model significantly underperforms for subgroups 2, 3, 4, and 6 compared to the OLS model. Next to this, the Ordered Logit model does not produce good results on individual level predictive performance.

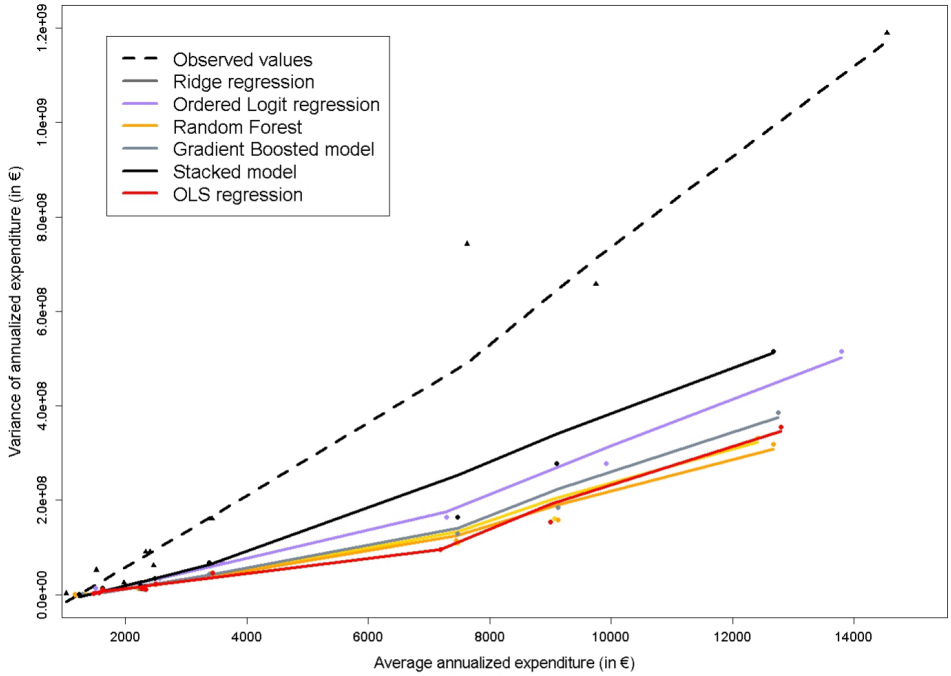

Figure 2: Mean annualized expenditure and variance of this expenditure within each subgroup.
Figure 2: Mean annualized somatic health care expenditure for each subgroup is displayed against the variance of annualized expenditure within each subgroup. Both the predictions by all models as observed values are plotted.

The main models of interest are the Random Forest, Gradient Boosted model and stacked model as these models outperform the OLS model on individual level. In terms of subgroup level, the Gradient Boosted model and stacked model display significantly lower predictive performance for subgroup 5, compared to the OLS model. However, for subgroups 2, 7 and 8 the predictive results are a significant improvement compared to the OLS model. The €107.27 overcompensation of subgroup 2 and undercompensations of €-437.93 and €-749.24 for subgroups 7 and 8 as estimated by OLS,

shrink with the use of the Random Forest, Gradient Boosted or stacked model. Especially the reduction in the undercompensation for subgroup 7 is large, as predicted annualized somatic health care expenditure by the Random Forest, Gradient Boosted and stacked model for this subgroup is approximately €300.00 higher. As subgroup 7 and 8 capture individuals that are historically undercompensated by a severe amount (>€5000), this result is promising.

In addition to the discussion of average under- and overcompensations for specific subgroups, it is interesting to note how each of the discussed models can capture observed variance within these subgroups. Figure 2 displays the relation between average annualized somatic health care expenditure and variance of this expenditure within the evaluated subgroups.

For both the predictions of each model and for the observed values, the mean and variance of annualized expenditure for each subgroup is plotted and a first-order local polynomial regression line is fitted through the points for each model and the observed values. Note that for each model as well as for the observed values, the relation between average annualized expenditure and the variance of annualized expenditure within subgroups is strictly positive.

Next to this, substantially higher variance is visible for the subgroups with high average annualized somatic health care expenditure. This result holds for both the observed values as for the predictions made by the models. The group of individuals captured in these subgroups therefore is considered much more heterogeneous in terms of annualized expenditure, compared to the subgroups 1 to 6 which capture on average low-cost individuals which display much lower variance in annualized somatic health care expenditure.

Note also that the variance of the predictions made by the models are substantially lower compared to the observed variance within the subgroups. This effect is especially visible for subgroups 7 to 9, which incorporate on average more high-cost individuals.

Finally, in the comparison of all models the stacked model best captures the variance within subgroups, as it is closest to the fitted line of observed variance in figure 2.

5 Discussion and conclusion

In this chapter, the results as presented in chapter 4 are discussed. Furthermore, the conclusions and the answer to the research question are presented, along with the limitations of this research and recommendations for further research.

5.1 Discussion

Based on the results displayed in table 10, the current Dutch OLS risk equalization model is not heavily outperformed by any of the alternative models on individual level. These results are in line with the literature, as in research performed by Stam et al. (Stam et al., 2020) on Dutch risk equalization data the OLS model was not heavily outperformed on individual level by machine learning techniques as well.

However, albeit just slightly, the outperformance of the OLS model by the Random Forest, Gradient Boosted and stacked model is visible on individual level. These models are slightly better able to predict annualized somatic health care expenditure on individual level compared to the OLS model. Highest predictive performance is visible for the stacked model. Therefore, for this dataset the historic claim that a combined predictive model is favored over any of the underlying single predictive models, as stated by Bates and Granger (1969), holds as well. This however does not present a clear motivation for the use of this model on the available risk equalization data, as the interpretability of the model is much lower compared to the OLS model.

On subgroup level, this research presents an insight in the possibilities for risk selection performed by health insurers. The models benchmarked against the current OLS method can be seen as models that could be exploited by Dutch health insurers, as the models use no more information than what is available to health insurers in practice. From that perspective, the predictive results for subgroups 2, 7 and 8 stand out. The Random Forest, Gradient Boosted and stacked model present significantly better results for these subgroups compared to the OLS model. Furthermore, for none of the evaluated subgroups the Random Forest is a worse fit to the data compared to the OLS model. For the Gradient Boosted and stacked model this holds as well, except for subgroup 5. For this subgroup, a higher undercompensation is predicted by these models compared to OLS.

Subgroup 2 captures individuals which are overcompensated in two of the previous three years. In the estimated OLS model, this overcompensation takes place again with an average magnitude of €107.27. This overcompensation is significantly reduced by the Random Forest, Gradient Boosted and stacked model as these models predict lower average annualized somatic health care expenditure compared to the OLS model. If one of these models is used by a health insurer, it presents a clear incentive for positive risk selection. The expected average compensation which is received based on the OLS estimation, is significantly higher than the expected average annualized expenditure within this subgroup as predicted by the Random Forest, Gradient Boosted or stacked model. Therefore,

individuals within this subgroup are expected to be overcompensated and thus attractive.

Subgroup 7 and 8 capture individuals who are severely undercompensated in the past three years. For subgroup 7 this is the result of one severe undercompensation in the past three years, for subgroup 8 this is the result of two undercompensations in the past three years. For both these subgroups, average annualized expenditure as predicted by the Random Forest, Gradient Boosted or stacked model is higher compared to the predictions made by the OLS model. Predictions for subgroups 7 and 8 by these models are therefore closer to the observed average annualized expenditure within these subgroups. This presents a clear incentive for negative risk selection on individuals in these subgroups. The expected average compensation as estimated by OLS is significantly lower compared to the average annualized expenditure as predicted by the Random Forest, Gradient Boosted model, and stacked model. Therefore, based on these predictions the individuals within these subgroups are on average loss-giving and are considered as unwanted from the perspective of an insurer.

The use of these models as benchmarks against the current OLS model therefore present clear possibilities to identify subgroups of individuals which are attractive or unattractive. The use of more detailed information about individuals within these subgroups can further identify the characteristics of these individuals. Note that this data was not available for this research but is available to health insurers.

Another remarkable result which is visible in table 11 is the indicative value of information from the past. This result motivates the use of historical information in the estimation process of risk equalization models. This is currently done by means of several variables which indicate if an insuree incurred high costs in the previous years. However, this information could possibly be exploited to a larger extent. For example, continuous health care expenditure data from the past could be used in the estimation process.

5.2 Conclusion

In the introduction to this research, the central research question was posed. It reads:

Which added value brings the application of the stacked algorithm to serve as a policy evaluation instrument for the current Dutch risk equalization model?

This question is answered by creating several alternative risk equalization models, using different techniques. Interpretability of these models is considerably lower compared to OLS, but this is not of high interest for the policy evaluation tool. For the evaluation, maximum predictive performance is of interest as this identifies possibilities for risk selection to the largest extent.

Presented with the predictive results of the models, an informed answer to the research question can be formed. From individual level perspective, the use of alternative models (which include the stacked model) to evaluate the current Dutch risk equalization model does not bring much added

value, as the model performance is only slightly better. However, on subgroup level the use of these alternative models presents meaningful insights. It proves that models with slightly better performance on individual level, are also able to better predict the annualized somatic health care expenditure for specific subgroups of the population, compared to the OLS model. This identifies clear opportunities for risk selection by health insurers. In addition, this research proves that the use of readily available information, such as historic results on risk equalization, to create subgroups can already bring specific subgroups to light which are expected to be under- or overcompensated by the OLS model compared to alternative model predictions. Health insurers have access to more detailed data, such as data underlying to the variables used in the risk equalization model. Therefore, health insurers are expected to be able to further identify the specific characteristics of these subgroups and adjust their (marketing) policy to these results.

5.3 Recommendations for further research

This research therefore provides a motivation to further research the predictive performance of risk equalization models on subgroup level, with the use of more detailed and preferably historic data. Next to this, it provides a motivation to benchmark the OLS model used in practice to new techniques, which can be exploited by health insurers.

Note, the results of this research should be interpreted in line with the limitations faced in this research. Given the confidentiality of the data, this research is performed in a restricted environment. This has limited this research in terms of the models used for estimation. The hyperparameter tuning process in chapter 4 is adjusted to the limited available computational power. Models that could present insightful results, such as tree-based algorithms with even more trees or the Ordered Logit model estimated via an iterative boosting procedure, were not feasible in terms of computational time and memory usage for this research. These are therefore models advised to evaluate further in future research.

Next to this, a limitation of this research is the unavailability of data underlying to the variables used in the actual risk equalization model. For this research, the exact same data set as used in practice for risk equalization is exploited. With the use of the data underlying to the variables in the risk equalization model, an even more sophisticated predictive model could be built or an even more precise characterization of the undercompensated people in the Dutch population could be made.

References

1. Bates, J., & Granger, C. (1969). The combination of forecasts. *Journal of Operational Research Society*, *20*(4), 451–468.

2. Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

3. Cumming, R., Knutson, D., Cameron, B., & Derrick, B. (2002). *A comparative analysis of claims-based methods of health risk assessment for commercial populations*. Society of Actuaries (USA).

4. Dutch Ministry of Health, W., & Sport. (2021). Regeling risicoverevening 2022. https://www.zorginstituutnederland.nl/publicaties/publicatie/2021/10/20/regeling-risicoverevening-2022

5. Ellis, R., Martins, B., & Rose, S. (2018). Risk adjustment for health plan payment. *Risk Adjustment, Risk Sharing and premium regulation in Health Insurance*, *1*, 55–104.

6. Ellis, R., & McGuire, T. (2007). Predictability and predictiveness in health care spending. *Journal of Health Economics*, *26*(1), 25–48.

7. ESHPM. (2018). WOR 930: Onderzoek Risicoverevening 2019 Berekening Normbedragen. https://www.zorginstituutnederland.nl/publicaties/publicatie/2018/10/11/wor-930-rapport-normbedragen-2019

8. ESHPM. (2019). WOR 974: Onderzoek Risicoverevening 2020 Berekening Normbedragen. https://www.zorginstituutnederland.nl/publicaties/publicatie/2019/10/30/wor-974-eindrapportage-normbedragen-2020

9. ESHPM. (2020). WOR 1002: Onderzoek Risicoverevening 2021 Berekening Normbedragen. https://www.zorginstituutnederland.nl/publicaties/publicatie/2020/10/16/wor-1002-eindrapportage-normbedragen-2021

10. Gilleskie, D., & Mroz, T. (2004). A flexible approach for estimating the effects of covariates on health expenditures. *Journal of Health Economics*, *23*(2), 391–418.

11. Gupta. (2020). Onderzoek Machine Learning in de Risicoverevening. https://www.rijksoverheid.nl/documenten/rapporten/2020/05/29/onderzoek-machine-learning-in-de-risicoverevening

12. Hastie, T., Tibshirani, R., & Friedman, J. (2017). *Elements of statistical learning: Data mining, inference and prediction*. Springer Series in Statistics.

13. Heij, C., de Boer, P., Franses, P. H., Kloek, T., & van Dijk, H. K. (2004). *Econometric methods with applications in business and economics*. Oxford University Press.

14. Jones, A., Lomas, J., Moore, P., & Rice, N. (2015). A quasi-monte-carlo comparison of parametric and semiparametric regression methods for heavy-tailed and non-normal data: An application to healhtcare costs. *Journal of the Royal Statistical Society: Series A*, *179*(4), 951–974.

15. Rose, S. (2016). A machine learning framework for plan payment risk adjustment. *Health Services Research*, *51*(6), 2358–2374.

16. Stam, P., Ismail, I., Visser, J., Portrait, F., & Koolman, X. (2020). Machine learning vereist betere risicoverevening zorgverzekeraars. https://equalis.nl/machine-learning-vereist-betere-risicoverevening-zorgverzekeraars/

17. Van Barneveld, E., Lamers, L., Van Vliet, R., & Van De Ven, W. (2000). Ignoring small predictable profits and losses: A new approach for measuring incentives for cream skimming. *Health Care Management Science*, *3*, 131–140.

18. Van Der Laan, M., Polley, E., & Hubbard, A. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1), 55–104.

19. Van Kleef, R., Eijkenaar, F., & Van Vliet, R. (2017). Risicoverevening 2016: Uitkomsten op de subgroepen uit de gezondheidsmonitor 2012. https://www.eur.nl/sites/corporate/files/Risicoverevening_2016_0.pdf

20. Van Kleef, R., Eijkenaar, F., & Van Vliet, R. (2020). Selection incentives for health insureres in the presence of sophisticated risk adjustment. *Medical Care Research and Review*, *77*(6), 584–595.

21. Van Kleef, R., Van Vliet, R., & Van De Ven, W. (2012). Risicoverevening tussen zorgverzekeraars: Kwantificering modelverbeteringen 1993-2011. *Tijdschrift voor Gezondheidswetenschappen*, (5), 312–326.

22. Van Kleef, R., Van Vliet, R., & Van De Ven, W. (2015). Innovatieve schattingsmethode voor risicoverevening. verkennend onderzoek naar mogelijkheden en effecten van constrained regression. https://www.eur.nl/sites/corporate/files/Eindrapport_Constrained_Regression_-_01jun15_0.pdf

23. Van Veen, S., Van Kleef, R., Van De Ven, W., & Van Vliet, W. (2015). Is there one measure-of-fit that fits all? A taxonomy and review of measures-of-fit for evaluating risk equalization models. *Medical Care Research and Review*, *72*(2), 220–243.

24. Van Veen, S., Van Kleef, R., Van De Ven, W., & Van Vliet, W. (2017). Exploring the predictive power of interaction terms in a sophisticated risk equalization model using regression trees. *Health Economics*, *27*(2), e1–e12.

25. Van Veen, S. (2016). *Evaluating and improving the predictive performance of risk equalization models in health insurance markets*. Erasmus University Rotterdam.

26. Vimont, A., Leleu, H., & Durand-Zaleski, I. (2022). Machine learning versus regression modelling in predicting individual healthcare costs from a representative sample of the nationwide claims database in france. *The European Journal of Health Economics*, *23*, 211–223.

27. Visser, J., Sonneveld, J., & Stam, P. (2014). *Het voorkomen van inadequate compensatie in de risicoverevening*. Strategies in Regulated Markets.

28. Wolpert, D. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259.

29. Zorgautoriteit, N. (2019). Monitor zorgverzekeringen. https://puc.overheid.nl/nza/doc/PUC_289640_22/1/

30. Zorgautoriteit, N. (2021). Monitor zorgverzekeringen. https://magazines.nza.nl/nza-magazines/2021/01/monitor-zorgverzekeringen-2021

31. Zorginstuut, N. (2021). Verantwoording Verzekerdenraming 2022. https://www.zorginstituutnederland.nl/financiering/publicaties/publicatie/2021/10/04/verantwoording-verzekerdenraming-2022

6 Appendix

6.1 Variables in the Dutch risk equalization models over time (2019-2022).

Table A1: Overview of variables and the number of categories used for the Dutch somatic ex-ante risk equalization models in the period 2019 to 2022.

Acronym	Description	2019	2020	2021	2022
Age/Gender	Interaction between age and gender	42	42	42	42
PCG	Pharmaceutical cost groups	38	38	39	43
DCG	Diagnostic cost groups	24*	24*	27	27
AMG	Medical equipment cost groups	11	11	15	15
SoI	Source of income	25	36	36	36
Region	Region based on postal address	10	10	10	10
SES	Social-economic status	12	12	12	12
PPA	Persons per address	13	13	13	13
MHC	Multiple-year high cost groups	9	9	9	9
PDG	Physiotherapy diagnosis cost groups	5	5	5	5
MNC	Multiple-year costs nursing and caring	10	10	10	10
HSM	Historical somatic morbidity	X	X	X	2
MPC	Multi-year pharmaceutical costs	X	X	X	2

Table A1: In the 2019 and 2020 risk equalization model, variable DCG was split up in two seperate variables: primary DCG and secondary DCG. In 2021 these are merged. In this table, the categories for primary and secondary DCG for 2019 and 2020 are added together, resulting in 24 categories.

6.2 Descriptive statistics for actual somatic health care expenditure in the full data set.

Table A2: Descriptive statistics of the actual Dutch somatic health care expenditure (in euros) for the period 2016 until 2019, for the full data set.

	2016	2017	2018	2019
Individuals	17,035,930	17,128,081	17,251,477	17,364,717
Insured years	16,749,741.44	16,839,944.32	16,949,009.52	17,058,676.05
Mean	2275.49	2278.76	2330.80	2407.93
St. Dev.	7699.61	7712.76	7865.29	8073.02
Skewness	19.36	17.90	19.40	18.34
Kurtosis	1549.24	983.59	1587.02	1094.11
Minimum	0.01	0.01	0.00	0.01
P1	42.51	40.66	30.74	4.81
Q25	156.99	158.23	166.79	176.81
Median	438.70	441.12	459.90	487.36
Q75	1,518.71	1511.12	1560.04	1635.98
P99	31,953.67	31,935.58	32,183.62	33,046.11
Maximum	2,495,408.86	1,493,545.72	2,617,403.93	1,911,534.72

Table A2: The number of insured years is the sum of the weights corresponding to all individuals. P1 represents the first percentile of actual health care expenditure distribution, Q25 and Q75 respectively the 25% and 75% quantiles and P99 represents the 99th percentile of actual health care expenditure distribution.

6.3 Descriptive statistics for annualized somatic health care expenditure in 2019.

Table A3: A detailed overview of the annualized somatic health care expenditure in 2019 (in euros), used for the 2022 risk equalization model.

	Somatic health care expenditure in 2019 (in euros)				
Population subset	Individuals	Insured years	Frequency	Mean	St. Dev
Age interacted with Gender					
Female (<65 years)	6,436,063	6,417,502.45	39.58%	1,820.52	6,544.03
Female (>65 years)	1,827,657	1,792,789.68	11.06%	5,745.85	12,029.05
Male (<65 years)	5,944,821	5,918,553.90	36.51%	1,293.06	6,703.53
Male (>65 years)	2,118,741	2,084,109.32	12.85%	5,348.32	13,145.83
Pharmacy based cost groups					
No indication (0)	13,368,220	13,301,986.74	82.05%	1,473.35	5,569.08
Indication (1-42)	2,959,062	2,910,968.61	17.95%	7,277.75	15,669.27
Diagnosis-based cost groups					
No indication (0)	14,314,192	14,241,696.52	87.84%	1,573.33	5,278.33
Indication (1-26)	2,013,090	1,971,258.84	12.16%	9,322.43	18,937.85
Medical-equipment based cost groups					
No indication (0)	15,523,937	15,434,618.22	95.20%	2,080.37	7,159.29
Indication (1-14)	803,345	778,337.13	4.80%	11,144.35	21,379.28
Source of Income interacted with age					
70+	2,409,074	2,348,620.21	14.49%	6,660.39	13,434.50
IVA benefit	154,339	153,278.02	0.95%	8,001.41	18,673.70
Incapacitated	930,844	926,858.98	5.72%	4,337.39	12,539.95
Assistance	640,794	638,097.08	3.94%	2,784.11	9,033.76
Student	593,370	595,111.15	3.67%	749.81	3,989.30
Self-employed	1,410,458	1,405,899	8.67%	1,260.27	5,709.67
High education	1,196,699	1,192,505.79	7.36%	1,197.57	4,653.95
Reference	8,988,704	8,952,584.36	55.22%	1,616.50	6,600.05
Region					
Region 1-5	8,093,602	8,028,452.66	49.52%	2,643.07	8,886.53
Region 6-10	8,223,680	8,184,502.70	50.48%	2,390.38	8,370.65
Socio-economic status interacted with age					
Very low	3,384,820	3,346,849.82	20.64%	3,208.78	9,723.58
Low	3,218,382	3,200,304.51	19.74%	2,661.06	8,837.78

Continued on next page

Table A3 : continued from previous page

Annualized somatic health care expenditure in 2019 (in euros)					
Population subset	Individuals	Insured years	Frequency	Mean	St. Dev
Middle	4,855,243	4,830,168.96	29.79%	2,349.15	8,238.89
High	4,868,837	4,835,632.06	29.83%	2,105.52	8,008.39
People per address interacted with age					
0-17 years	2,833,745	2,828,521.11	17.45%	1,105.41	6,668.12
WLZ long-term stay	197,981	177,773.28	1.10%	3,224.83	8,869.97
WLZ influx	42,915	42,842.50	0.26%	17,789.25	22,968.88
1-person household	2,282,882	2,257,354.01	13.92%	4,248.48	11,261.48
Reference	10,969,759	10,906,464.45	67.27%	2,450.97	8,217.99
Multiple-year high cost groups					
No indication	8,727,778	8,693,197.89	53.62%	794.92	3,585.19
In one year	6,519,381	6,473,108.72	39.93%	2,948.46	7,838.89
In two years	169,698	164,754.15	1.02%	10,608.96	19,766.61
In 3 years	910,425	881,894.60	5.44%	14,786.24	22,921.41
Physiotherapy diagnosis-based cost groups					
No indication (0)	15,875,368	15,768,244.05	97.26%	2,338.59	8,070.22
Indication (1-4)	451,914	444,711.31	2.74%	8,788.66	19,132.27
Multiple-year costs for nursing and caring					
No indication (0)	15,893,214	15,803,327.76	97.47%	2,212.93	7,345.36
Indication (1-9)	434,068	409,627.70	2.53%	17,799.55	25,105.94
Historical somatic morbidity					
No indication (0)	8,375,165	8,339,075.59	51.43%	1,086.40	5,020.31
Indication (1)	7,952,117	7,873,879.76	48.57%	4,029.05	11,056.18
Multiple-year pharmaceutical cost groups					
No indication (0)	11,023,593	10,979,503.28	67.72%	1,052.76	4,461.93
Indication (1)	5,303,689	5,233,452.08	32.28%	5,584.29	13,232.50

Table A3: A subset of this table is presented in the main text, in table 2. The number of insured years is the sum of the weights corresponding to all individuals. The frequency is calculated based on the relative number of insured years. The presented mean and standard deviation correspond to annualized somatic health care expenditure in 2019.

6.4 Descriptive statistics for the subsets of data used in this research.

Table A4: Descriptive statistics of annualized somatic health care expenditure in 2019.

	Learning set	Cross-Val. set	Training set	Validation set	Test set
Individuals	13,060,528	2,000,000	10,450,146	2,610,382	3,266,754
Insured years	12,968,874	1,986,001	10,377,096	2,591,778	3,244,075
Mean	2514,31	2518.72	2514.06	2515.31	2520.31
Median	506,16	507.04	506.29	505.65	506.92
St. dev.	8161,05	8293.58	8151.61	8157.60	8255.25

Table A4: Descriptive statistics for the five different subsets of the full data sample (16,327,282 observations), used in this research. Mean, median and the standard deviation of annualized expenditure is displayed in euros.

6.5 Predictive results: unscaled expenditure level

Table A5: Unscaled individual level predictive performance of the evaluated models for the test set.

Model		MSE	R^2	CPM
Base models	Ridge regression	47,821,631	0.368	0.351
	Ordered Logit	55,240,734	0.270	0.358
	Random Forest	47,298,957	0.375	0.358
	Gradient Boosted model	47,246,117	0.376	0.357
Stacked model		46,875,094	0.381	0.352
OLS regression		47,321,583	0.375	0.386

Table A5: Unscaled individual-level predictive performance for the evaluated models.

Table A6: Unscaled subgroup level predictive performance in terms of MPE (in euros) of the evaluated models for the test set.

Group	Observed	Ridge	Ord. Logit	RF	GBM	Stacked	OLS
1	1523,93	228.83***	24.31***	210.08***	206.98***	242.50***	62.96
2	2406,67	255.77***	211.71***	255.73***	242.53***	293.68***	41.02
3	3430,59	189.17***	236.50***	187.03***	170.78***	236.88***	-87.85
4	2335,69	119.71***	-0.12***	142.31***	132.62***	180.45***	-56.39
5	2469,93	-112.31***	-151.90***	-62.21***	-85.21***	-34.89***	-253.95
6	1981,34	-370.82***	-447.77	-302.71***	-311.51***	-279.37***	-458.12
7	7626,01	338.68***	-187.23***	322.60***	328.37***	469.97***	-627.34
8	9751,06	-81.58***	361.14***	-22.35***	-39.52***	116.40***	-986.45
9	14542,22	-1306.80***	-473.07***	-1024.44***	-977.22***	-804.06***	-2081.61

Table A6: Observed mean annualized expenditure is displayed for each subgroup. The unscaled average deviation from this mean, denoted with the MPE, is presented for all models. MPE values indicated with (***), (**) or (*) differ significantly from MPEvalues obtained by the OLS model on respectively 1%, 5% or 10% significance level as tested by means of a two-sample t-test.

6.6 Final model estimation results.

6.6.1 Model estimation results: parametric regressions

Table A7 displays the estimated regression coefficients for the OLS model, the Ordered Logit regression with K=10 intervals and the Ridge regression with $\lambda = 254$, all trained on the full learning set. Both the Ridge regression and Ordered Logit regression capture intercepts and therefore need to omit categories of variables in the estimation procedure. For each variable in the data, one of the categories is therefore omitted from the estimation model. For the variable Age interacted with gender, for both genders one category is withheld from the estimation procedure. For variables interacted with age, a category is withheld from the estimation procedure for each different age group.

Table A7: Estimation results for the OLS, Ordered Logit and Ridge regression model.

	OLS	Ord. Logit	Ridge
Age interacted with Gender			
1-4 years (M)	2398.04***	-	-
5-9 years (M)	2093.80***	0.02**	-305.88
10-14 years (M)	2079.09***	−0.22***	-341.65
15-19 years (M)	2173.07***	−0.04***	-251.60
20-24 years (M)	1964.49***	0.03***	-473.67
25-29 years (M)	1963.20***	0.07***	-488.90
30-34 years (M)	1968.31***	0.16***	-474.91
35-39 years (M)	2012.36***	0.13***	-472.61
40-44 years (M)	2060.48***	0.22***	-428.39
45-49 years (M)	2159.53***	0.28***	-347.35
50-54 years (M)	2273.95***	0.38***	-228.48
55-59 years (M)	2502.07***	0.55***	37.88
60-64 years (M)	2704.90***	0.74***	264.86
65-69 years (M)	2969.08***	0.86***	437.80
70-74 years (M)	3344.11***	0.95***	1062.53
75-79 years (M)	3710.67***	1.13***	1641.84
80-84 years (M)	4067.83***	1.23***	2365.70
85-89 years (M)	4572.13***	1.57***	3323.87
90+ years (M)	5453.13***	1.99***	4749.02
1-4 years (F)	2109.92***	-	-
5-9 years (F)	2077.42***	−0.06***	-326.18

Continued on next page

Table A7 : continued from previous page

	OLS	Ord. Logit	Ridge
10-14 years (F)	2122.10***	−0.13***	-294.24
15-19 years (F)	2282.75***	0.32***	-140.14
20-24 years (F)	2224.88***	0.46***	-276.80
25-29 years (F)	2663.80***	1.10***	193.15
30-34 years (F)	2813.92***	1.13***	349.83
35-39 years (F)	2419.62***	0.68***	-62.99
40-44 years (F)	2219.10***	0.39***	-273.91
45-49 years (F)	2260.78***	0.50***	-245.83
50-54 years (F)	2309.84***	0.53***	-207.21
55-59 years (F)	2381.5***	0.53***	-108.75
60-64 years (F)	2515.45***	0.67***	47.07
65-69 years (F)	2727.34***	0.80***	181.57
70-74 years (F)	2852.58***	0.89***	620.56
75-79 years (F)	3217.25***	1.03***	1025.36
80-84 years (F)	3688.68***	1.27***	1638.62
85-89 years (F)	4223.95***	1.55***	2329.08
90+ years (F)	4856.18***	1.87***	3420.57
Pharmacy based cost groups			
Residual group	-286.56***	-	-
Thyroid affection	-59.15**	0.10***	142.79
Glaucoma	31.18	0.17***	148.05
Depression	−6.03	0.11***	273.09
Psychosis	39.13	0.26***	412.73
Epilepsy	343.98***	0.23***	582.71
Chronic anticoagulation	633.31***	0.31***	772.26
Transplants	2338.81***	1.11***	4567.06
Parkinson	2781.71***	0.55***	2954.51
Heart diseases: other	1679.94***	0.37***	2457.35
Chronic pain excl. opioids	731.70***	0.33***	988.72
Neuropathic pain	1261.23***	0.22***	1412.89
Diabetes II without hypertension	253.45***	0.29***	519.84
Diabetes II with hypertension	638.78***	0.34***	840.32
Diabetes I without hypertension	1102.49***	0.89***	1318.73
Diabetes I with hypertension	1681.88***	0.71***	1902.28
Cystic fibrosis/Pancreatic enzymes	3162.00***	0.54***	3472.04

Continued on next page

Table A7 : continued from previous page

	OLS	Ord. Logit	Ridge
Growth disorders (add-on)	2300.93***	2.75***	2325.59
Brain disorder / spinal: other	3297.68***	0.42***	3339.46
Brain disorder / spinal: MS	4563.61***	2.15***	4521.72
HIV/AIDS	1077.18***	1.29***	869.69
Psoriasis	356.17***	0.44***	494.61
Crohns disease / Ulcerative colitis	411.42***	0.46***	554.94
Rheumatishm	589.33***	0.34***	751.05
Autoimmune diseases (add-on)	2311.48***	1.54***	2603.74
Kidney diseases	8644.96***	0.64***	10060.25
Acromegaly	14447.35***	2.21***	151013.77
Immunoglobulin (add-on)	12532.93***	0.66***	11675.34
Asthma	150.95***	0.26***	388,02
COPD / severe asthma	1278.94***	0.55***	1936.64
COPD / Severe asthma (add-on)	11645.76***	2.04***	11754.49
Hormone-sensitive tumors	819.10***	0.26***	1212.48
Cancer	696.21***	0.54***	593.02
Cancer (add-on)	8402.97***	1.06***	9461.48
Pulmonary arterial hypertension	15458.83***	3.47***	14633.08
Macular degeneration	2287.62***	0.66***	2719.05
Hypercholesterolemia	2467.55***	1.62***	2345.77
Heart diseases: anti-arrhytmics	576.13***	0.30***	607.07
Addiction excl. nicotine	1052.79***	0.40***	1450.60
Extreme high costs cluster 1	101324.15***	15.69***	98668.08
Extreme high costs cluster 2	180774.57***	11.78***	183717.92
Extreme high costs cluster 3	327867.71***	18.75***	317929.20
Extreme high costs cluster 4	487160.17***	8.19***	486621.43
Diagnosis-based cost groups			
Residual group	−371.40***	-	-
DCG 1	189.21**	0.25***	392.38
DCG 2	763.30***	0.31***	1178.92
DCG 3	1014.34***	0.41***	1357.13
DCG 4	1691.38***	0.56***	2421.47
DCG 5	2309.74***	0.43***	3193.74
DCG 6	2922.30***	0.54***	3896.22
DCG 7	2969.54***	0.83***	4357.03

Continued on next page

Table A7 : continued from previous page

	OLS	Ord. Logit	Ridge
DCG 8	3466.81***	0.68***	5596.09
DCG 9	4113.30***	0.84***	5959.23
DCG 10	4234.28***	1.20***	6422.42
DCG 11	4636.97***	0.43***	5136.81
DCG 12	5444.85***	0.74***	7216.38
DCG 13	5397.0***2	0.70***	8844.55
DCG 14	7884.01***	0.97***	10422.30
DCG 15	7696.00***	1.03***	8180.67
DCG 16	10865.20***	1.49***	15086.89
DCG 17	13089.04***	1.11***	13429.64
DCG 18	11600.71***	1.62***	12591.57
DCG 19	10938.59***	0.17***	10130.49
DCG 20	14203.68***	1.16***	22504.45
DCG 21	14088.11***	1.89***	17384.69
DCG 22	17891.62***	1.42***	22032.82
DCG 23	20573.71***	2.24***	21458.49
DCG 24	30061.94***	2.72***	38498.96
DCG 25	50876.98***	4.01***	52618.76
DCG 26	55435.22***	24.18***	56780.23
Medical equipment based cost groups			
Residual group	−77.37***	-	-
MCG 1	376.82***	0.35***	318.33
MCG 2	292.79***	0.16***	247.14
MCG 3	1448.75***	0.50***	1316.20
MCG 4	3158.64***	0.16***	3460.52
MCG 5	1920.32***	0.51***	3227.51
MCG 6	1830.22***	0.30***	2804.03
MCG 7	3233.02***	0.38***	5876.22
MCG 8	7430.32***	0.70***	9829.83
MCG 9	13020.85***	1.70***	10997.19
MCG 10	6694.59***	1.00***	9394.00
MCG 11	1868.47***	0.20***	2755.88
MCG 12	999.60***	0.40***	771.49
MCG 13	1972.31***	0.83***	1945.45
MCG 14	1063.21***	0.75***	1048.30

Continued on next page

Table A7 : continued from previous page

	OLS	Ord. Logit	Ridge
Source of Income interacted with age			
70+ years	0	-	-
IVA benefit 0-17 years	53.00	0.22***	34.65
IVA benefit 18-34 years	1070.74***	0.54***	710.38
IVA benefit 35-44 years	916.14***	0.73***	497.18
IVA benefit 45-54 years	774.30***	0.45***	302.97
IVA benefit 55-64 years	658.62***	0.42***	77.83
IVA benefit 65-69 years	402.46***	0.23***	613.23
Incapacitated 0-17 years	102.98***	0.25***	66.84
Incapacitated 18-34 years	303.34***	0.25***	118.19
Incapacitated 35-44 years	471.02***	0.43***	281.49
Incapacitated 45-54 years	414.60***	0.42***	219.77
Incapacitated 55-64 years	323.20***	0.37***	-1.64
Incapacitated 65-69 years	390.99***	0.23***	526.50
Social assistance 0-17 years	113.34***	0.07***	82.75
Social assistance 18-34 years	191.93***	0.11***	131.54
Social assistance 35-44 years	225.38***	0.20***	164.81
Social assistance 45-54 years	256.63***	0.17***	181.51
Social assistance 55-64 years	253.42***	0.29***	42.11
Social assistance 65-69 years	251.78***	0.21***	519.09
Student 0-17 years	23.36	0.16***	-108.31
Student 18-34 years	−159.07***	−0.29***	-194.89
Self-employed 0-17 years	−53.33**	−0.09***	-97.69
Self-employed 18-34 years	−53.69*	−0.12***	-106.28
Self-employed 35-44 years	−102.94***	−0.15***	-112.45
Self-employed 45-54 years	−125.18***	−0.20***	-105.44
Self-employed 55-64 years	−175.33***	−0.06***	-161.38
Self-employed 65-69 years	−38.81	−0.02	140.90
Higher education 0-17 years	−93.10**	−0.15***	-155.85
Higher education 18-34 years	9.83	−0.00	1.57
Higher education 35-44 years	−57.01***	−0.13***	-63.03
Reference group 0-17 years	−3.19***	-	-
Reference group 18-34 years	22.95***	-	-
Reference group 35-44 years	−20.97***	-	-
Reference group 45-54 years	−45.48***	-	-

Continued on next page

Table A7 : continued from previous page

	OLS	Ord. Logit	Ridge
Reference group 55-64 years	-63.24^{***}	-	-
Reference group 65-69 years	-100.76^{***}	-	-
Region			
Region 1	39.77^{***}	0.05^{***}	126.90
Region 2	32.86^{***}	0.07^{***}	110.11
Region 3	14.07	0.03^{***}	87.40
Region 4	11.64	0.05^{***}	82.45
Region 5	-3.19	-0.01^{**}	55.05
Region 6	-5.95	0.01^{*}	51.24
Region 7	-15.60	-0.07^{***}	18.03
Region 8	-13.52	0.02^{***}	33.11
Region 9	-18.09	0.04^{***}	36.38
Region 10	-39.08^{***}	-	-
Socio-economic status interacted with age			
Very low 0-17 years	28.92	0.44^{***}	-35.66
Very low 18-69 years	-30.03	-0.06^{***}	-35.36
Very low 70+ years	-144.52^{***}	0.04^{***}	-1152.90
Low 0-17 years	15.62	0.31^{***}	-71.47
Low 18-69 years	15.12^{*}	-0.06^{***}	31.66
Low 70+ years	-0.67	0.03^{***}	-514.78
Middle 0-17 years	-14.97	0.22^{***}	-93.45
Middle 18-69 years	23.45^{***}	-0.01^{***}	40.39
Middle 70+ years	59.57^{***}	0.07^{***}	-90.73
High 0-17 years	-13.97	-	-
High 18-69 years	-13.58^{***}	-	-
High 70+ years	67.83^{***}	-	-
People per address interacted with age			
Children 0-17 years	0	-	-
Wlz-institution long-term 18-69 years	-658.25^{***}	-0.42^{***}	-351.87
Wlz-institution long-term 70-79 years	-2102.62^{***}	-0.80^{***}	-1251.22
Wlz-institution long-term 80+ years	-3171.51^{***}	-1.46^{***}	-2375.25
Wlz-institution influx 18-69 years	10416.74^{***}	0.41^{***}	9787.16
Wlz-institution influx 70-79 years	$11303.85v$	1.32^{***}	11211.20
Wlz-institution influx 80+ years	8904.14^{***}	1.57^{***}	8252.64
One-person household 18-69 years	16.22	0.08^{***}	40.92

Continued on next page

Table A7 : continued from previous page

	OLS	Ord. Logit	Ridge
One-person household 70-79 years	235.82***	0.09***	755.97
One-person household 80+ years	226.88***	0.04***	667.01
Residual group 18-69 years	−4.28***	-	-
Residual group 70-79 years	−121.30***	-	-
Residual group 80+ years	−193.86***	-	-
Multiple-year high cost groups			
Residual group	−531.51***	-	-
Top 30% costs in at least 1 of last 3 years	62.28***	1.22***	538.94
Top 10% costs in the 2 preceding years	2211.84***	1.92***	2656.73
Top 15% costs in the 3 preceding years	1994.53***	1.93***	2332.19
Top 10% costs in the 3 preceding years	3277.83***	2.23***	3523.34
Top 7% costs in the 3 preceding years	4891.83***	2.41***	5290.16
Top 4% costs in the 3 preceding years	8468.73***	2.67***	8515.81
Top 1.5% costs in the 3 preceding years	17986.79***	3.20***	17194.59
Top 0.5% costs in the 3 preceding years	43739.72***	3.31***	41801.38
Physiotherapy diagnosis-based cost groups			
Residual group	-23.70***	-	-
PDG 1	404.54***	0.55***	345.82
PDG 2	1285.73***	0.49***	1221.50
PDG 3	5800.22***	0.16**	7492.82
PDG 4	9485.96***	0.98***	8673.75
Multiple-year costs for nursing and caring			
Residual group	−174.92***	-	-
Top 3.5% costs in 3 preceding years	1009.37***	0.31***	1724.16
Top 3.0% costs in 3 preceding years	1520.56***	0.35***	2257.59
Top 2.5% costs in 3 preceding years	2763.12***	0.76***	4007.17
Top 2.0% costs in 3 preceding years	5100.00***	1.07***	6747.56
Top 1.5% costs in 3 preceding years	7661.44***	1.59***	9511.28
Top 1.0% costs in 3 preceding years	10949.06***	2.13***	12815.54
Top 0.5% costs in 3 preceding years	15317.72***	2.28***	17068.16
Top 0.25% costs in 3 preceding years	26153.80***	2.72***	28413.15
Top 0.25% costs in 3 preceding years (0-17 years)	56544.37***	5.21***	55202.38
Historical somatic morbidity			
Residual group	-95.36***	-	-
HSM 1	101.00***	0.40***	202.19

Continued on next page

Table A7 : continued from previous page

	OLS	Ord. Logit	Ridge
Multiple-year pharmaceutical cost groups			
Residual group	-151.41***	-	-
MPC 1	317.63***	0.64***	514.18
Intercept			
Intercept	-	-	769.48
Intercept 1—2	-	2.93***	-
Intercept 2—3	-	4.15***	-
Intercept 3—4	-	5.12***	-
Intercept 4—5	-	5.95***	-
Intercept 5—6	-	6.74***	-
Intercept 6—7	-	7.63***	-
Intercept 7—8	-	8.52***	-
Intercept 8—9	-	9.57***	-
Intercept 9—10	-	11.28***	-

Table A7: The coefficients for the OLS, Ordered Logit and Ridge regression estimated on the full learning set. For the OLS and Ordered Logit regression, coefficients indicated with an (***), (**) or (*) are significantly different from 0 at 1%, 5% or 10% significance level.

6.6.2 Model estimation results: tree-based regression models

Table A8 displays variable importance for the tree-based non-parametric regression models used in this research. For the Random Forest, impurity is displayed. This is measured as the increase in squared errors of the predictions obtained by a model fitted on randomly permuted values of a specific variable, compared to the model fitted on actual values for this variable. This thus denotes the predictive importance of each specific variable. The variable importance values of the Random Forest presented in table A8 are divided by 1,000,000.

Shap values are presented for the Gradient Boosted model. These values indicate the mean absolute marginal contribution of each variable to the predictions made for the test set by this model.

For both importance measures it holds that higher values in table A8 indicate higher predictive importance of variables.

Table A8: Variable importance for the tree-based Random Forest and Gradient Boosted model.

	Random Forest	Gradient Boosted model
Age interacted with gender		
1-4 years (M)	341,392	3.26
5-9 years (M)	49,894	0.19
10-14 years (M)	89,712	0.72
15-19 years (M)	72,920	0.04
20-24 years (M)	49,911	30.82
25-29 years (M)	75,039	22.68
30-34 years (M)	27,080	18.80
35-39 years (M)	163,250	18.10
40-44 years (M)	41,744	12.22
45-49 years (M)	99,917	6.53
50-54 years (M)	77,260	0.01
55-59 years (M)	118,066	8.46
60-64 years (M)	250,556	19.73
65-69 years (M)	168,064	29.40
70-74 years (M)	148,840	0.59
75-79 years (M)	152,184	8.71
80-84 years (M)	148,993	1.88
85-89 years (M)	155,064	6.97
90+ years (M)	163,612	7.88
1-4 years (F)	36,480	0.36
5-9 years (F)	25,158	0.70
10-14 years (F)	61,302	0.56
15-19 years (F)	13,637	1.94
20-24 years (F)	108,816	0.39
25-29 years (F)	33,665	22.90
30-34 years (F)	36,977	30.85
35-39 years (F)	36,075	8.01
40-44 years (F)	48,441	0.03
45-49 years (F)	62,450	0.14
50-54 years (F)	59,774	1.30
55-59 years (F)	175,806	0.04
60-64 years (F)	156,398	5.59
65-69 years (F)	82,355	17.27

Continued on next page

Table A8 : continued from previous page

	Random Forest	Gradient Boosted model
70-74 years (F)	144,278	15.09
75-79 years (F)	107,868	1.15
80-84 years (F)	101,891	5.03
85-89 years (F)	88,029	1.34
90+ years (F)	126,331	11.61
Pharmacy based cost groups		
Residual group	2,243,619	50.32
Thyroid affection	159,378	1.77
Glaucoma	224,445	1.67
Depression	324,513	0.92
Psychosis	90,004	0.72
Epilepsy	161,939	1.50
Chronic anticoagulation	132,708	4.61
Transplants	215,135	4.52
Parkinson	154,872	4.01
Heart diseases: other	823,917	46.97
Chronic pain excl. opioids	263,116	6.38
Neuropathic pain	168,635	4.41
Diabetes II without hypertension	107,293	1.17
Diabetes II with hypertension	157,914	7.27
Diabetes I without hypertension	268,515	3.97
Diabetes I with hypertension	269,160	13.65
Cystic fibrosis/Pancreatic enzymes	340,559	1.77
Growth disorders (add-on)	40,970	0.22
Brain disorder / spinal: other	50,736	1.20
Brain disorder / spinal: MS	20,511	2.14
HIV/AIDS	20,121	1.20
Psoriasis	84,544	0.10
Crohns disease / Ulcerative colitis	103,094	1.01
Rheumatishm	82,516	0.65
Autoimmune diseases (add-on)	89,053	5.37
Kidney diseases	160,209	4.05
Acromegaly	39,658	2.16
Immunoglobulin (add-on)	167,515	2.35
Asthma	221,248	2.35

Continued on next page

Table A8 : continued from previous page

	Random Forest	Gradient Boosted model
COPD / severe asthma	244,308	29.71
COPD / Severe asthma (add-on)	24,438	1.54
Hormone-sensitive tumors	132,392	3.20
Cancer	9,359	0.07
Cancer (add-on)	2,464,087	27.69
Pulmonary arterial hypertension	68,396	1.31
Macular degeneration	46,034	4.86
Hypercholesterolemia	11,777	1.07
Heart diseases: anti-arrhytmics	54,253	0.11
Addiction excl. nicotine	67,349	1.10
Extreme high costs cluster 1	410,550	2.24
Extreme high costs cluster 2	316,953	1.58
Extreme high costs cluster 3	1,094,220	1.66
Extreme high costs cluster 4	524,086	0.55
Diagnosis-based cost groups		
Residual group	4,066,249	156.30
DCG 1	364,652	8.98
DCG 2	310,786	23.48
DCG 3	501,906	35.53
DCG 4	844,441	25.61
DCG 5	841,932	59.22
DCG 6	430,043	28.82
DCG 7	347,042	21.96
DCG 8	217,621	6.88
DCG 9	249,187	8.42
DCG 10	256,484	6.48
DCG 11	65,981	2.74
DCG 12	2,708,034	44.05
DCG 13	131,753	1.67
DCG 14	412,484	9.56
DCG 15	355,958	7.42
DCG 16	659,173	22.99
DCG 17	324,692	2.86
DCG 18	129,773	6.05
DCG 19	330,240	3.35

Continued on next page

Table A8 : continued from previous page

	Random Forest	Gradient Boosted model
DCG 20	375,206	6.14
DCG 21	156,730	3.99
DCG 22	476,387	5.34
DCG 23	538,570	12.97
DCG 24	894,394	7.12
DCG 25	2,536,382	20.42
DCG 26	149,852	2.09
Medical equipment based cost groups		
Residual group	1,519,754	57.23
MCG 1	90,547	6.46
MCG 2	108,726	10.29
MCG 3	91,905	1.50
MCG 4	155,998	2.49
MCG 5	410,084	11.02
MCG 6	217,923	3.52
MCG 7	198,116	6.06
MCG 8	311,659	3.97
MCG 9	42,610	0.30
MCG 10	256,559	4.65
MCG 11	165,135	8.60
MCG 12	112,160	0.66
MCG 13	33,620	0.58
MCG 14	15,565	0.15
Source of Income interacted with age		
70+ years	823,976	221.34
IVA benefit 0-17 years	11,430	0.00
IVA benefit 18-34 years	13,252	0.05
IVA benefit 35-44 years	19,505	0.35
IVA benefit 45-54 years	53,750	0.51
IVA benefit 55-64 years	62,044	1.08
IVA benefit 65-69 years	46,697	0.62
Incapacitated 0-17 years	28,259	0.33
Incapacitated 18-34 years	53,780	2.70
Incapacitated 35-44 years	41,903	2.24
Incapacitated 45-54 years	49,419	3.15

Continued on next page

Table A8 : continued from previous page

	Random Forest	Gradient Boosted model
Incapacitated 55-64 years	71,255	3.72
Incapacitated 65-69 years	95,937	5.32
Social assistance 0-17 years	26,537	0.45
Social assistance 18-34 years	17,721	0.00
Social assistance 35-44 years	12,038	0.39
Social assistance 45-54 years	29,029	1.53
Social assistance 55-64 years	52,398	2.41
Social assistance 65-69 years	54,940	1.87
Student 0-17 years	7,688	0.00
Student 18-34 years	17,235	18.57
Self-employed 0-17 years	62,355	1.57
Self-employed 18-34 years	6,937	3.22
Self-employed 35-44 years	13,873	5.04
Self-employed 45-54 years	43,438	2.44
Self-employed 55-64 years	72,941	0.06
Self-employed 65-69 years	58,181	0.09
Higher education 0-17 years	65,361	0.43
Higher education 18-34 years	36,934	1.37
Higher education 35-44 years	32,960	3.74
Reference group 0-17 years	120,576	1.40
Reference group 18-34 years	59,389	4.00
Reference group 35-44 years	135,537	0.09
Reference group 45-54 years	112,034	2.37
Reference group 55-64 years	299,898	34.08
Reference group 65-69 years	139,730	0.00
Region		
Region 1	354,924	8.88
Region 2	243,239	5.00
Region 3	224,184	1.81
Region 4	256,472	2.53
Region 5	261,245	0.06
Region 6	206,668	0.00
Region 7	272,202	2.98
Region 8	183,543	1.17
Region 9	255,640	0.06

Continued on next page

Table A8 : continued from previous page

	Random Forest	Gradient Boosted model
Region 10	224,589	9.83
Socio-economic status interacted with age		
Very low 0-17 years	36,847	0.06
Very low 18-69 years	92,808	3.51
Very low 70+ years	141,425	13.43
Low 0-17 years	249,233	0.17
Low 18-69 years	131,244	2.62
Low 70+ years	108,107	2.01
Middle 0-17 years	72,209	0.75
Middle 18-69 years	121,856	6.97
Middle 70+ years	198,715	9.32
High 0-17 years	74,841	1.10
High 18-69 years	302,327	1.12
High 70+ years	268,604	10.16
People per address interacted with age		
Children 0-17 years	141,493	17.13
Wlz-institution long-term 18-69 years	40,924	3.64
1 Wlz-institution long-term 70-79 years	20,957	12.35
Wlz-institution long-term 80+ years	54,519	44.34
Wlz-institution influx 18-69 years	169,838	8.52
Wlz-institution influx 70-79 years	137,603	6.63
Wlz-institution influx 80+ years	226,295	15.97
One-person household 18-69 years	156,515	11.11
One-person household 70-79 years	152,575	3.07
One-person household 80+ years	213,089	20.20
Residual group 18-69 years	218,227	7.63
10 Residual group 70-79 years	141,863	12.16
Residual group 80+ years	161,285	17.25
12 **Multiple-year high cost groups**		
Residual group	1,216,257	357.44
Top 30% costs in at least 1 of last 3 years	1,991,908	82.87
Top 10% costs in the 2 preceding years	134,598	18.39
Top 15% costs in the 3 preceding years	307,264	34.41
Top 10% costs in the 3 preceding years	216,286	31.89
Top 7% costs in the 3 preceding years	493,804	41.28

Continued on next page

Table A8 : continued from previous page

	Random Forest	Gradient Boosted model
Top 4% costs in the 3 preceding years	1,042,846	52.60
Top 1.5% costs in the 3 preceding years	1,022,521	20.93
Top 0.5% costs in the 3 preceding years	4,599,873	24.13
Physiotherapy diagnosis-based cost groups		
Residual group	164,743	8.85
PDG 1	128,541	0.51
PDG 2	200,152	11.47
PDG 3	293,456	0.54
PDG 4	18,136	0.04
Multiple-year costs for nursing and caring		
Residual group	5,701,667	269.45
Top 3.5% costs in 3 preceding years	174,905	30.23
Top 3.0% costs in 3 preceding years	258,884	24.74
Top 2.5% costs in 3 preceding years	92,579	17.41
Top 2.0% costs in 3 preceding years	110,836	7.02
Top 1.5% costs in 3 preceding years	191,680	3.08
Top 1.0% costs in 3 preceding years	372,526	15.15
Top 0.5% costs in 3 preceding years	288,398	13.54
Top 0.25% costs in 3 preceding years	1,556,847	24.64
Top 0.25% costs in 3 preceding years (0-17 years)	225,086	2.43
Historical somatic morbidity		
Residual group	20,821	1.32
HSM 1	433,757	101.67
Multiple-year pharmaceutical cost groups		
Residual group	32,592	7.21
MPC 1	1,528,869	172.80

Table A8: Variable importance for the tree-based regression models used in this research. For the Random Forest, variable importance is calculated as the increase in squared errors of the predictions obtained by a model fitted on randomly permuted values of a specific variable, compared to the model fitted on actual values for this variable. For the Gradient Boosted model, shap values are displayed which equal the average absolute marginal contribution of a variable to predictions made by the Gradient Boosted model.