2-8-2023

Applying machine learning methods to the Menzis district nursing benchmark

J.P. Eugelink
Faculty of Science and Technology
Master Health Sciences

Internal supervisors

First supervisor: dr. M.M.A Kip

Second supervisor: dr. S. Saing

External supervisors

First supervisor: M. Tenniglo Msc

Second supervisor: dr. A.P. Hayen

Abstract

Background: health insurer Menzis wants to build a new benchmark model with which the performance of providers of district nursing care can be compared to each other. It will be used for monitoring the performance of providers, and to assist during contract negotiations. The benchmark model predicts district nursing costs based on the characteristics of clients. These expected costs are then compared to the realised costs. The goal of this thesis is to explore different machine learning methods and assess their suitability for use by Menzis.

Method: firstly, a number of variable selection and data manipulation steps are taken. After this, linear regression, elastic net regression, regression trees, random forests and propensity score matching are applied to the Menzis benchmark. These techniques are then assessed based on three dimensions: the variable selection, the fit of the models, and the usability of the models. The last dimension will be made measurable by gathering input from health care purchasers, who are the intended users of the benchmark model.

Results: demographic variables such as age, gender and income, and health status variables such as use of district nursing care in the year prior and medication use in the year prior were found to be among the most important predictors of district nursing costs. Linear regression, elastic net regression and random forests all show similar predictive value. Health care purchasers have a strong preference for the use of propensity score matching, with random forests as the second favourite.

Discussion and conclusion: propensity score matching and random forests seem to be the most suitable methods for building a district nursing benchmark model. For future development of the benchmark model, it is recommended that the variable selection and data manipulation steps are thoroughly reconsidered.

Table of contents

Introduction

Benchmarking concerns the comparison of performance measures between competitors [1]. Analysis of competitors or organisations that serve a similar purpose can help to provide insight into the practices and performance of other organisations, or into what has already been tried to solve specific problems. This way, plans to address the need for improvement can be developed, and successful ideas from others can be borrowed and adapted.

Ideally, benchmarking works in cycles of making comparisons, discovering best practices, and lastly implementing these best practices [2]. This means it is part of the process of continuous quality improvement [1]. To draw a valid comparison between competitors, the practices of each competitor have to be mapped first: this is the first step while carrying out benchmarking. Thus, a large advantage of benchmarking is the fact that it forces people to look beyond their own practices, and helps to identify opportunities for improvement [3]. Another useful by-product of benchmarking can be found even before that, as the process that is going to be benchmarked will have to be fully mapped beforehand. It was found that this initial mapping of the process at hand often improved the understanding of the process for everyone involved [1]. This means that benchmarking not only forces people to look beyond their own process, but also forces them to investigate their own processes first. Lastly, benchmarking encourages people to think about what constitutes ‘good’ performance, and how to make it measurable.

The focus of this thesis will be on formal benchmarking. Formal benchmarking is characterized by the use of a systematic approach, which means that steps such as data collection and variable selection are carefully considered [1]. This is necessary, because the goal of this thesis is to develop a benchmark model that assists during contract negotiations with health care providers. This means that the benchmark model could make a big impact on providers and Menzis itself. Ways to make a benchmark more systematic include setting a clear goal for the benchmark, and carefully choosing which performance measures to include. The other type of benchmarking is called informal benchmarking, which involves very little to no planning. It is often defined as ‘industrial tourism’.

Furthermore, this thesis will focus on external benchmarking, as the benchmark that is examined aims to compare performance measures across different organisations [4]. If different departments within the same organisation would be benchmarked against each other, it would be called internal benchmarking [2]. The choice between internal and external benchmarking is a part of the important step of setting the scope of the benchmark.

In a health care context, benchmarking is often defined as ‘comparing performance measures across providers’ [5]. However, when taking a quality management approach, the focus lies more on the opportunity to learn from the experience of others. This means that the data that has been collected to compare different people, departments or institutions can be used as a learning tool. This changes the definition to ‘using someone else’s successful process as a measure of desired achievement for the activity at hand’ [3]. The key change in this definition is the addition of learning from others. It can be concluded that learning from others is an important positive by-product of benchmarking.

One of the more well-known examples of benchmarking in health care is the benchmark that was developed by the Dutch Health Care Authority (NZa) [6]. This benchmark was developed to monitor trends in the productivity of hospitals in the Netherlands, as well as their cost-effectiveness. The information that is gathered from the benchmark helps the NZa in performing its core tasks: monitoring health care institutions and insurers, and advising the ministry of Health, Welfare and Sport [7].

Aside from making and enacting policy, benchmarking in the health care sector can also be performed by health insurance companies. Health insurers use benchmarking to determine if an institution will be contracted and on what terms [8]. During this decision-making process, insight into the performance of these health care institutions can be a powerful tool for the health insurer. Benchmarking can also be used by health insurers to monitor the performance of providers while they are contracted. This information can be used to find discrepancies in performance, and to help providers make adjustments to their performance. Examples of performance measures that could be used in this case are the amount of complications that occur, and the amount of care that is delivered.

One of the health insurers that makes use of such benchmarking is Menzis [9]. Menzis is one of the largest health insurers in the Netherlands, with roughly 1.8 million policy holders [10]. The company uses various benchmarking models to compare different health care providers. One of these benchmark models concerns district nursing. District nursing encompasses all health care that is delivered by nurses in the patient's own environment [11]. This can be at home, at work or anywhere that is not a health care institution. District nursing is covered under the standard insurance package in the Netherlands.

The goal of using this benchmark model is to monitor the performance of the providers of district nursing care. The benchmark model predicts the average use of care (in euros) per policy holder regarding district nursing for each health care provider [9]. The model bases its predictions on the characteristics of the policy holders, such as, age, gender, socioeconomic status and education. The outcome of the benchmark is the difference between the predicted cost and the realised cost. Based on the results of the benchmark, Menzis can look for an explanation for any discrepancies. For example, it is likely that there is a positive correlation between age and district nursing utilisation [12]. Because of this, the benchmark model will predict a higher cost for patients with a higher age. This way, the predicted costs for a health care provider with an older patient population will be higher. As a result of including age in the benchmark model, providers will not get punished for having an older patient population, which is something they cannot influence. This way, a health care provider's high cost could be explained by the average age of the patients it services. The information generated by the benchmark model can then be used during contract negotiations, or to encourage a health care provider to improve its performance. This is the 'learning' part of the benchmark, which is, as mentioned before, a vital part of benchmarking.

Menzis currently uses a large dataset with characteristics of the policy holders to make the predictions used in the benchmark. The goal of this thesis is to critically evaluate the district nursing benchmarking model Menzis currently uses, and to explore the possibilities of using other machine learning methods to improve the model. This will be done by identifying possible methods that could be used, and then assessing their fit to the case at hand. This leads to the following research question: which machine learning method is the best fit for predicting use of care in the case of policy holders from Menzis that receive district nursing? The following dimensions will be used to assess the improvement made compared to the original benchmark model by Menzis: the variable selection of the model, the fit of the model, and lastly the usability of the model for analysts and health care purchasers at Menzis.

Theoretical framework

Literature review

An important step of benchmarking is choosing which variables to include in the benchmark model. The variable selection for the models presented in this thesis was done by means of a literature review. A search was conducted to determine whether various demographic and health status variables influence the health care utilization and/or costs regarding district nursing care of an individual. When no information about a connection between the variable of interest and district nursing utilization could be found, the search was broadened to include other similar forms of care administered at home. Finally, when this did not provide any results, the search was broadened again to include all forms of health care.

The search was carried out between the 15th and 25th of April 2023, using the search engine Scopus. The only additional inclusion criterion was that articles must be written in the English or Dutch language. Here, an example of a query that was used is listed: (“socioeconomic status” OR income”) AND (“health care use” OR “health care utilization” OR “health care usage” OR “health care cost*”. The part between the first set of parentheses was adjusted to search for each variable and its synonyms.

Based on this literature review, the following variables were found to be (potentially) related to district nursing utilization:

- Research by Linden et al. found that the family composition of a patient strongly predicts the utilization of caregiving services [13]. People that live alone are more likely to use increased levels of caregiving services, while the opposite is the case for people with children. This is because people that live in the same household as someone that needs caregiving can provide informal care.
- According to data from Statistics Netherlands (CBS), people with a higher age are more likely to develop one or more chronic diseases [14]. Additionally, Duncan et al. found that health care costs increase significantly during the final year of life [15].
- A publication by Kempen et al. reported that the income of a patient affects the utilization of home care among the elderly in the Netherlands [16]. Data from CBS shows that education level influences the health status of a patient: the lower the education level, the higher the prevalence of chronic disease [14]. This in turn results in higher health care utilization, as found by De Meijer et al [17]. This same research found that gender is a strong predictor for long term care utilization in the Netherlands.
- Research by Wingen and Otten found that the socioeconomic status of a patient serves as a predictor for their physical and mental health, as well as the amount of disabilities experienced in daily life [18].
- A publication by Andersen and Newman suggests utilization of health care services is not only influenced by demographic and social structural variables, but also attitudinal-belief variables [19].

Which variables should be included in a benchmark?

Now that information about which variables have an effect on the outcome of interest has been obtained, choices have to be made about which variables to include in the benchmark model, and in what capacity they should be included. Not all variables that have an effect on the outcome of interest should be put into the benchmark model.

The most important type of variable that should not be put into a benchmark model without consideration is a variable that can be influenced by the person or organisation that is being benchmarked. For example, when benchmarking sprinters against each other, the outcome of interest could be the time it takes the sprinter to run 100 meters. It is most likely that a strong relationship between time spent in training and the outcome of interest is found. However, including this variable in the benchmark might not be appropriate. When this variable is corrected for in the benchmark, a sprinter who spent little time in training will be expected to set a slower 100 meter time. This means that the sprinter is not 'punished' by the benchmark for spending too little time in training. The opposite is also true: a sprinter that spends a lot of time in training will not be 'rewarded' by the benchmark. Thus, including training time in the benchmark masks performance. One solution for this is to perform the benchmark twice: once where the variable that can be influenced is excluded, and once where it is included. This way, the sprinter that did not train enough will stand out as running slower than expected when the training time is not adjusted for in the benchmark. When the benchmark is performed again, but with training time included, it is likely that the same sprinter does not perform worse than expected according to this benchmark. It can then be concluded that this sprinter's underperformance can be (partly) explained by them not training enough.

When benchmarking, variables can be corrected for through either concurrent or prospective adjustment [20]. Concurrent adjustment uses information from the performance year of the benchmark. Prospective adjustment makes use of information prior to the start of the performance year.

Another solution for influenceable variables is to use prospective adjustment for those variables. For example, using concurrent adjustment for health status variables might lower the incentive to prevent disease, as health care providers could raise their expected costs by making more diagnoses than they should [21]. Using prospective adjustment for these variables resolves this problem: when health status variables from a previous year are used, a provider that inflates their diagnoses will now be punished with a lower benchmark score. This means that there is an incentive for providers to keep their patient population as healthy as possible.

Machine learning methods

In the following section, various machine learning methods will be described. Machine learning allows computers to 'learn' without being explicitly programmed to do so [22]. Here, 'learning' can be defined as the computer being able to discern patterns in a dataset. These patterns can then be used to make predictions about new, unknown data points.

First, some methods that can be used to predict outcomes for new data points will be discussed. After this, three regularization methods that can be used to optimize the model will be explained. At a later stage, the suitability of these methods for the benchmark model will be assessed.

Linear regression

Linear regression is a technique that tries to describe the relationship between variables as a linear function [23]. This can be done using a single variable (simple linear regression) or multiple variables (multiple linear regression). The method can also be used to make predictions about future data. In simple linear regression, the equation for a straight line is used: $y = ax + e$. Here, the dependent variable is commonly continuous. For multiple linear regression, the formula changes to: $y = a_1x_1 + a_2x_2 + e$ [24]. This method will fit a straight line to the data in the way that results in the smallest sum of squared residuals. This is called the Ordinary Least Squares (OLS) method. The following assumptions are inherent to the linear regression model: there is some form of linear relationship between the predictor(s) and the outcome variable (linearity) [23,25]. The variation around the regression line is roughly the same for each value of x (homoskedasticity). This variation around the regression line follows a normal distribution for each value of x (normality). Lastly, the deviation of each individual data point from the regression line is independent of that of other data points (independence).

Logistic regression

When the dependent variable is not continuous, linear regression is not the best choice, because then it violates the assumption that the variation around the regression line is roughly the same for each value of x: thus, it introduces heteroskedasticity. In the case of a dichotomous dependent variable, logistic regression is a suitable method [25]. It can describe the relationship of both continuous and discrete independent variables with the dependent variable. Similarly to linear regression, logistic regression can also use a single independent variable (simple logistic regression) and multiple independent variables (multiple logistic regression). The coefficients that follow from a logistic regression are expressed as odds ratios (OR) to ease interpretation. In the case of a dichotomic independent variable, this is calculated by dividing the odds of exposure in cases where the event of interest occurs by the odds of exposure in cases where the event of interest does not occur [26]. An example when predicting costs for a patient using gender as the independent variable: an OR of 2 would mean that the odds of having costs above a certain cut-off point for patients that are male are twice as high than the odds for patients that are female. As logistic regression is only capable of using dichotomous dependent variables, a cut-off point has to be chosen when the dependent variable of interest is continuous.

Regression tree

In a situation where a linear relationship between the predictor(s) and the outcome variable cannot be assumed, a regression tree could be a good choice. This is because regression trees are able to accommodate a wide range of (non-linear) relationships. They are also able to capture interactions between variables of interest. A regression tree is a decision tree where the response variable is continuous [27]. This is the case for the Menzis benchmark, as the outcome measure here is district nursing costs. The structure of a regression tree can be seen in figure 1. Regression trees consist of a

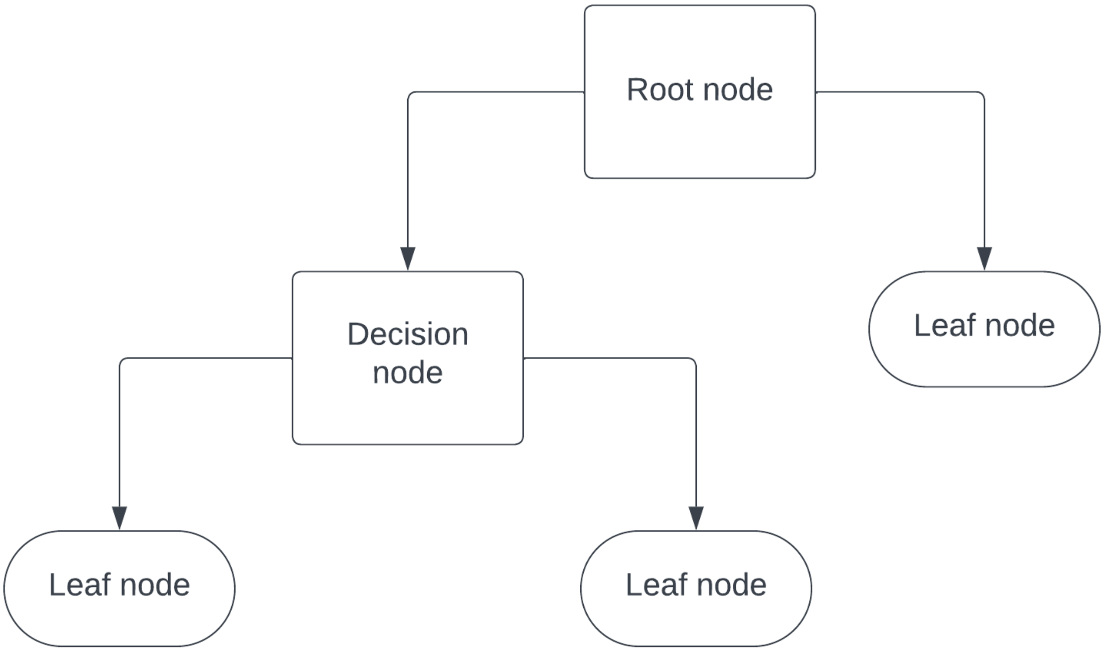


Figure 1: the structure of a regression tree

root node, representing the first split of the data. Every next node where a split is made is called a decision node. A node where no further split is made, is called a leaf node. Regression trees can make use of both categorical (e.g. gender) and numeric (e.g. income) explanatory variables. Splits for categorical variables are made using the levels of the variable. Splits for numeric variables are defined by choosing a value, then dividing the data points based on whether their value is smaller or greater than the chosen value. The splits are selected based on the homogeneity within the resulting two nodes: the higher the homogeneity, the better. The homogeneity of nodes is determined by calculating the impurity. This can be calculated in various ways. One example is the Gini Index, which is defined as $GI = 1 - \Sigma c^2$. Here, c is the proportion of responses in each category. The split with the highest homogeneity, and thus the lowest impurity, is chosen as the root node. When a new data point appears, the data point is run through all the nodes in the tree. The leaf node where it ends up constitutes the prediction of the regression tree for that particular data point. A key advantage of regression trees is that they are easy to visualize and interpret [28]. A regression tree with multiple variables can easily be interpreted, whereas plotting results in different ways often gets harder and harder to interpret when using more than two variables. When using a regression tree, the size of the tree should be carefully chosen: a tree that is too small will result in inaccurate predictions, while a tree that is too large will be overfitted. When building regression trees in R, the size of the tree is determined by the complexity parameter (CP). When this parameter equals 0, the tree size is not limited, and the program will find the tree with the best fit. When the value of this parameter rises, the size of the tree decreases. This is useful to reduce the amount of overfitting, and to keep trees small enough to be interpretable.

Random forests

Random forests can be seen as the creation of multiple regression trees at once [29]. This is useful, because regression trees tend to work well with the data that is used to train them, but when it comes to making predictions for new data points, their performance strongly decreases. The trees that make up a random forest differ from each other because of multiple factors. Firstly, for each tree, a random subset of the training data is chosen to construct the tree. This process is called ‘bagging’, as the data are first bootstrapped, and the aggregate is used to make a prediction. Secondly, a random subset of the available variables is chosen. This randomness reduces overfitting, thus circumventing a problem that the singular regression tree suffers from. Random forests work by letting multiple trees ‘vote’ on a prediction. For a classification task, all trees make a prediction, and the class with the most ‘votes’ is chosen as the prediction. For regression tasks, the prediction of each tree is averaged, which leads to the final prediction.

Propensity score matching

The propensity score is defined as ‘the probability of treatment assignment conditional on observed baseline covariates’ [30]. This means that subjects with similar characteristics will have a similar propensity score [31]. Propensity scores are commonly calculated using logistic regression. After each subject is assigned a propensity score, treated subjects can be matched to untreated subjects with a similar propensity score. This ensures that each treated subject will be compared to an untreated subject with similar characteristics. This way, the effect of treatment can be determined more accurately, as differences in the outcome variable between a treated and untreated person with a similar propensity will not be due to differences in observed patient characteristics. Matching can be done in various ways: for example, a subject from the treatment group can be matched to a subject from the other group whose propensity score is closest to the propensity score of the treated subject (nearest neighbour matching). Furthermore, subjects can be matched one-to-one or one-to-many. For the latter, one treated subject is matched do multiple untreated subjects. This form of matching can be useful to increase the size of the matched sample, and thus reducing the variance in the estimation of treatment effect [32,33]. It does, however, introduce bias as treated subjects will be matched to untreated subjects with less similar propensity scores. Thus, a trade-off between bias and variance is in place. A condition to getting matches of sufficient quality is a region of common support between the treated group and the untreated group [33]. This means that there is an overlap in the propensity scores of the treated and control subjects. If there is no region of common support, treated subjects will be matched to untreated subjects with dissimilar propensity scores, which equates to low quality matches. To further evaluate the quality of matching, the balance of matching should be evaluated. Matching is balanced when treated subjects and the untreated subjects they are matched to have similar distributions of measured characteristics [34]. Propensity score matching is often used to mimic a randomised controlled trial in situations where performing one is not feasible. This could be due to the costs associated with performing a randomised controlled trial.

Regularization

Regularization is a concept that can be used to prevent overfitting of a model. Overfitting occurs when the model provides accurate predictions for the training data, but not for new data. Regularization can be used when there is a high number of variables, and a small number of observations [35,36]. Regularization is especially useful when it is likely that these variables are correlated with each other. In the case of OLS regression, this would result in incorrect parameter estimates for those that are correlated. Regularization works by shrinking the coefficient estimates of variables that are correlated with each other. This is done by utilizing some form of penalty. Using an OLS regression as an example: usually, the model picks the line that results in the smallest sum of squared residuals. When applying regularization, the model will try to minimize the sum of squared residuals with a penalty added to it. These penalties come in different forms. Below, three ways of applying penalties are described.

Ridge regression works by adding the ridge penalty to the sum of squared residuals. The ridge penalty is defined as $\lambda * a^2$ [36]. Here, a is defined as the coefficient of the independent variable. Instead of fitting a straight line to the data while minimizing the sum of squared residuals, the model will minimize the sum of squared residuals with the ridge penalty added to it. In practice, this means ridge regression will shrink the coefficients of predictors that are correlated with each other. At higher levels of λ , the coefficients will be shrunk more. This means the model will become less sensitive to the predictors that contribute largely to the sum of squared residuals.

Lasso regression is similar to ridge regression, as it is also a regularization method that shrinks the coefficients of variables [35,37]. However, lasso regression is also capable of performing variable selection. This is because lasso regression can shrink a coefficient all the way to 0, whereas ridge regression can only shrink a coefficient asymptotically close to 0. The fact that lasso regression is able to shrink a coefficient to 0 improves the interpretability of the resulting model, as fewer variables will be used to predict the outcome. This also makes the model easier to use in practice, as there are fewer variables that need data collection. Lasso regression adds a penalty to the sum of the squared residuals that is defined as $\lambda * |a|$.

The regularization method that will be applied in this thesis is elastic net regression [38]. This method is useful when the amount of variables is very high, and the usefulness of the variables is uncertain. Elastic net regression applies a combination of the ridge penalty and lasso penalty. It tries to minimize the sum of squared residuals with the following penalty added to it: $\lambda * ((\alpha * |a|) + (1 - \alpha) * \mathrm{a}^2)$. Like before, a stands for the coefficient of the variable the penalty is applied to. As can be seen in the formula, the weights of the two different penalties are determined by the value of α. When α = 0, the lasso penalty turns to zero, which means only the ridge penalty is applied. When α = 1, the ridge penalty turns to zero, which means only the lasso penalty is applied. When the value for α lies between 0 and 1, a mixture of both penalties is applied. The best values for both α and λ are found by using cross-validation. Elastic net regression combines the capability of ridge regression to shrink multiple coefficients at once with the capability of lasso regression to eliminate coefficients entirely.

Performance measures

There are various ways of evaluating model performance. All performance measures have in common that they aim to give an indication of how well the model can predict the outcome of interest. Each performance measure has their own advantages and drawbacks, which is why various performance measures will be applied in this thesis.

The first of these is the Mean Absolute Error (MAE), which is the average of the absolute difference between the predicted values and the actual values. A lower MAE indicates a better fit of the model to the data. Next, the Root Mean Square Error (RMSE) is calculated. This is done by taking the square root of the average of the squared difference between the predicted values and the true values. The RMSE is more sensitive to outliers than the MAE, because it squares the deviations first [39]. This means that, by looking at the difference between the MAE and RMSE value, conclusions can be drawn about the distribution of errors. For example, when the value for RMSE is higher than that of the MAE, it can be concluded that this is because there are instances where the error is large. The larger the value of RMSE compared to MAE, the more large errors there are. Lastly, the R^2 will be calculated. The value of R^2 indicates what percentage of variance in the outcome is explained by the model. A higher R^2 translates to a higher predictive value of the model. Research by Chicco at al. suggests that R^2 is more informative and easier to interpret than measures like the MAE and RMSE mentioned before [39].

Method

The starting data

The benchmark will be built using two different datasets, which contain characteristics about all 2.2 million clients that were policyholders at Menzis in either the calendar year 2021 or 2022. The characteristics can be divided into two groups: demographics, such as age and family composition, and variables regarding health care resource utilization, such as the amount of days of district nursing received and the type of care received. In total, 118 variables are included in these datasets. Using R, the two datasets were merged based on the client number, which stays the same for each individual Menzis client over the years. This results in a dataset where only clients that were policyholders in both years are included. A code that references the health care provider of the client is included to be able to assign the costs of a client to the correct provider. Since the datasets contain the same variables, all duplicate columns were removed, except for the district nursing cost in 2021. This means the merged dataset now contains values for both district nursing cost in 2021 and 2022.

To be able to draw valid comparisons between the existing model and the new models, some exclusion criteria have to be applied. All clients under the age of 18 were filtered out, which is also the case for the existing benchmark model. This choice was made because health care for children and teenagers strongly differs from health care for adults. For the same reason, clients that died during 2022 were excluded, as they were not clients for the full year. Lastly, clients that used more than one provider of district nursing care were excluded, as it cannot be determined which part of the cost should be attributed to a specific provider.

Missing data were found in the variables regarding the type of care the client receives. Since the amount of missing data is small, all rows with missing data were removed from the dataset. This resulted in 1384 rows being excluded. For the demographic variables, no missing data were found.

Data manipulation

Variables were included in the analysis based on the literature review that was presented in the theoretical framework: a variable is included when at least one article was found that suggests that the variable influences health care costs and/or utilization. In this section, it will be explained which variables are included in the analysis, and in what capacity. Any manipulation of variables is reported in the paragraphs below.

- The number of people under and over the age of 18 are included in the analysis, as well as an indicator whether there are two earners in the household. These variables were chosen as they give the most information about the family composition of a client out of the various variables concerning family composition.
- Age was added as a categorical variable, to be able to capture a non-linear relationship between age and district nursing costs. Age categories spanning 5 years were chosen, as these are the categories that the Menzis analysts commonly use in their analyses.
- As health care costs increase in the final year of life, a variable that indicates whether a client died in the first 6 months of 2023 is included.
- Income is included as a categorical variable. The levels of the categorical variables can be found in table 1.
- Education level is included in as a categorical variable.
- Gender is included as a binary variable.
- The Menzis dataset contains a variable that is based on the attitude of the client: whether they are performance-oriented, docile, self-willed, etc. The origin, validity and objectivity of this

variable remain questionable. However, it is included because of the use of regularization methods later on, which will remove the variable from the analysis if it is not useful.

- A variable that indicates whether a client received district nursing in the previous year was added. This is a form of prospective adjustment.

Lastly, 63 variables regarding the health care utilization of the client are included. These consist of variables that indicate the health status of the patient. More specifically, diagnoses-based cost groups (diagnose kosten groepen) give information about the chronic diseases a patient has, based on diagnoses that were set during a hospital admission in the past [40]. Pharmacy-based cost groups (farmaceutische kosten groepen) are groups where clients are categorised based on the prolonged use of certain medication [41]. Assistive tool cost groups (hulpmiddelen kosten groepen) give information about the use of assistive tools in the past. These variables can be used as proxies for health status, as the cost groups serve as markers for chronic conditions [41]. The cost group variables originate from 2021, so they are a form of prospective adjustment.

Variables regarding the amount of care received, such as the number of days that care was received, were dropped, as the outcome measure of interest is the cost of district nursing. Another reason for dropping these variables is that the different types of district nursing care use various types of time registration. Because of this, it is not possible to simply add up the amount of care delivered to multiple patients. Variables regarding the location of the client, such as the postal code, were dropped, as Menzis wants to compare providers across the whole country. Adjusting for geographical location would mean that providers are compared with other providers from their region, which could cover up underperformance when multiple providers in a region underperform. Lastly, variables regarding the treatment codes of a client were excluded, as these variables do not provide any information that the cost group variables mentioned above do not.

The sample

The filters and variable selection described above lead to a sample which contains 43917 clients and 79 variables. Clients in this dataset are between 18 and 104 years old, with an average age of 75. On average, they used €5167 worth of district nursing care in 2022. Out of the sample, 38% is male, and 62% is female. There are some clients with a large number of people in the same household, with numbers as high as 49. This could be explained by them living in some form of assisted living facility. The full summary statistics of this sample are provided in table 1, which is listed below. The rightmost column in table 1 indicates the reference variable on which the inclusion of the variable in the analysis is based. For example, diagnoses-based cost groups are included because a relationship between health status and health care costs was found during the literature review.

N = 43917	**Mean**	**Standard Deviation**	**Median**	**Min**	**Max**	**Reference variable**
Age	75	12	78	18	104	Age [14]
District nursing cost 2021 (€)	4054	7256	453	0	119816	Health status [17]
District nursing cost 2022 (€)	5167	7475	1908	5	103367	
Number of people under the age of 18 in household	0.04	0.29	0	0	8	Family composition[13]
Number of people over the age of 18 in household	1.5	1.02	1	1	49	Family composition [13]

	Frequency	Proportion	
Gender			Gender [17]
Male	16676	0.38	
Female	27241	0.62	
Education level			Education level [14,18]
Primary education	11388	0.26	
Lbo/vmbo (crafts)/mbo 1	12510	0.29	
Mavo/mulo/vmbo (theoretical)	4253	0.09	
Mbo 2, 3 or 4	10540	0.24	
Havo/vwo/hbs	1241	0.03	
Hbo or academic bachelor	3190	0.07	
Hbo or academic master/PhD	1241	0.02	
Income			Income [18][16]
<€18000	5083	0.12	
€18000-€26000	10455	0.24	
€26000-€35000	7049	0.16	
€35000-€50000	8037	0.18	
€50000-€75000	6997	0.16	
€75000-€100000	3559	0.08	
>€100000	2737	0.06	
Two earners in household			Family composition [13]
No partner and/or not classified	370007	0.84	
Yes	3288	0.07	
No	3622	0.08	
Deceased between 01-01-2023 and 01-06-2023			Increased healthcare cost in final year of life [15]
Yes	2175	0.05	
No	41742	0.95	
Has used specialist medical care in 2021			Health status [17]
Yes	41323	0.94	
No	2594	0.06	
Type of client			Attitudinal-belief variables [19]
Consumption oriented	9913	0.23	
Quality oriented	2414	0.06	
Result oriented	3341	0.08	
Luxury oriented	3906	0.09	
Convenience oriented	5270	0.12	
Socially critical	2669	0.06	
Docile	15139	0.35	
Wayward	1265	0.03	

Diagnoses-based cost groups			
No cost group	186	0.004	Health status [17]
1 cost group	37816	0.86	
More than 1 cost group	5915	0.13	
Pharmacy-based cost groups			Health status [17]
1 cost group	32614	0.74	
More than 1 cost group	11303	0.26	
Assistive tool cost groups			
No cost group	1413	0.03	
1 cost group	40567	0.92	Health status [17]
More than 1 cost group	1937	0.05	

Table 1: summary statistics of sample

Data analysis

All data analyses are carried out using R. The methods described in the theoretical framework will be applied to the Menzis case. The choice was made to use elastic net regression as the only regularization method, as this method combines the advantages of both ridge and lasso regression. Also, elastic net regression is also capable of fully functioning like a ridge or lasso regression when this turns out to be optimal. Firstly, the dataset is split up into 4 different datasets. The division is made based on the number of months a client has received district nursing care. This results in 5 different tables that are used in all analyses: one table with all clients in it, and one table each for the groups with 1 to 3 months, 4 to 6 months, 7 to 9 months and 10 to 12 months of district nursing care. This division is made to be able to see if the predictive value of the models differs based on whether a client received short term or long term care. Each of these 5 datasets were split into training and testing sets. A train/test split of 0.7/0.3 was chosen. Because some diagnostic and pharmaceutical cost groups are very uncommon, the train/test split resulted in variables with no variance. These had to be removed, as the models in R cannot function when there are variables with no variance. Thus, 3 variables concerning medication use were removed.

Differences between the Menzis benchmark model and the new models

The existing district nursing benchmark model was built using the same dataset as the new benchmark models that are created in this thesis. In the following section, the differences between the existing model and the new models will be described. Any differences regarding variable selection or data manipulation will be listed. Only differences between the models will be described here: when the existing model and the new models match on a particular part, it will not be described in this section. The following bullet points concern variables that were present in the existing model, but were not included in the new models presented in this thesis.

- The variables that indicate the number of people in the household of the client above and below the age of 18 were merged into one variable that indicates the number of people in the household. Household sizes were capped at 5 people: when there are 5 or more people in the household, the value for this variable will be 5.
- A new variable, the sum of pharmacy-based cost groups (FKGs), was added. This variable indicates how many FKGs the client has, by counting how many FKG variables have a 1 in

the dataset. A maximum value of 5 FKGs was chosen: when there are 5 or more FKGs, the value for the sum of FKGs will be 5.

- Socio-economic class was added as a categorical variable. This variable is a function of the income and education level variables. The levels of this variable are ‘A’, ‘B1’, ‘B2’, ‘C’, and ‘D’, in descending order of socio-economic class. This variable is a function of two variables that are also included in the analysis: income and education level.
- An interaction variable ‘Gender * Age’ was added, where the age variable to the power of 6 is multiplied by the gender variable (female = 1). This was done to capture the interaction between these variables.

The existing benchmark model was originally made in 2020, and also uses data from that year. To be able to draw a valid comparison between the existing model and the new models, the former was adapted to run on data from 2022. For the existing model, the data was also split into training and testing set, again to ensure a valid comparison. This was not done when the model was built. Linear regression was used to predict the district nursing cost of clients. In table 2 below, the reasoning for deviating from the existing Menzis benchmark model will be addressed.

Menzis model	**New models**	**Reasoning**
Members in household as 1 variable	Members of household below and above the age of 18 separately	This variable was split into two to be able to capture the difference in the capability of children and adults to provide informal care.
Sum of pharmacy-based cost groups added as a variable	Sum of pharmacy-based cost groups not added as a variable	The sum of pharmacy-based cost groups was not added, because one cost group contains multiple medications. Some cost groups contain more than others. Thus, a sum of the cost groups is not necessarily an indicator of the amount of different medications used by a client.
Socioeconomic class variable added	No socioeconomic class variable added	Due to time constraints, interaction variables were not considered for the new models presented in this thesis.
Gender * age^6 added	No interaction variable added	Idem
No indicator for multiple people with a job in household added	Indicator for multiple people with a job in household added	This variable was added because it provides additional information on family composition.
No indicator for district nursing use in past year added	Indicator for district nursing use in past year added	This variable was included as a proxy for health status.
People above the age of 100 excluded	People above the age of 100 included	People above the age of 100 are included, because no evidence was found that the district nursing care process strongly differs for this group.

All deceased clients excluded	Clients that died in the 6 months after the benchmark year included	Including clients that died within 6 months after 2022 makes it possible to correct for end of life care that is supplied through district nursing.
4% highest cost for each age category excluded	No outliers removed	Outliers were not removed for the new models, as the data did not suggest that the highest amounts were wrongly calculated or entered. Furthermore, the clients with the highest amounts of costs are spread across a large number of providers. As can be seen in appendix 1, The removal of the 4% of clients with the highest costs would result in clients with costs above €25,000 being removed.
No train/test split used	Train/test split of 0.70/0.30	Splitting the data into training and testing sets is important when building the model. This is to check how the model performs when predicting on data it has not seen before. This is important, because the predictive value on new data is ultimately what is of interest.

Table 2: the reasoning behind deviating from the existing benchmark model

Elastic net regression

As mentioned in the theoretical framework, when using elastic net regression, values for α and λ have to be chosen. Firstly, the value for α is determined by using a loop that tries all values of α between 0.0 and 1.0. The value that results in the lowest mean squared error is chosen as the value for α. After this, the value for λ is determined via 10-fold cross validation. If possible, the value for λ that results in in the model with the fewest parameters and a mean squared error within 1 standard error of the λ that results in the lowest mean squared error is chosen. However, this will sometimes result in a model where all parameters are shrunk to 0. In this case, the λ that results in the lowest mean squared error is chosen. Because the R package that is used for elastic net regression is not capable of handling categorical variables directly, all categorical variables were transformed into dummies.

Regression tree

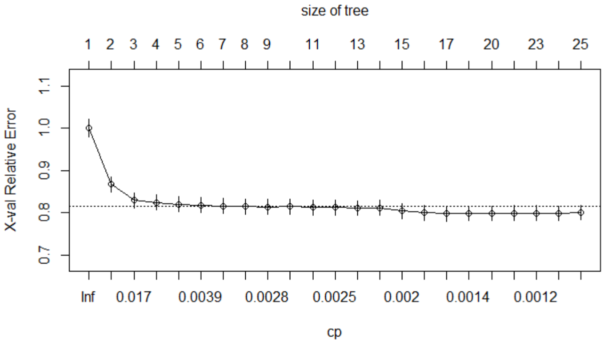


Figure 2: example of complexity parameter plot

For the regression trees, a value for the complexity parameter (CP) has to be chosen. The amount of end nodes a tree has is determined by the value of CP. The value for CP was found by creating a tree that is overfitted, which happens when the CP is 0. This way, the tree is not restricted in how many splits it produces. After this, the results of using different values for CP can be plotted, as seen in figure 2 [42]. The leftmost value of CP that results in a relative error below the dotted horizontal line is chosen. This is done separately for each tree. In the case of this example, the optimal amount of end nodes for this tree would be 15, as this is the leftmost point where the graph is below the dotted line. The corresponding CP value for this would be 0.002.

Random forest

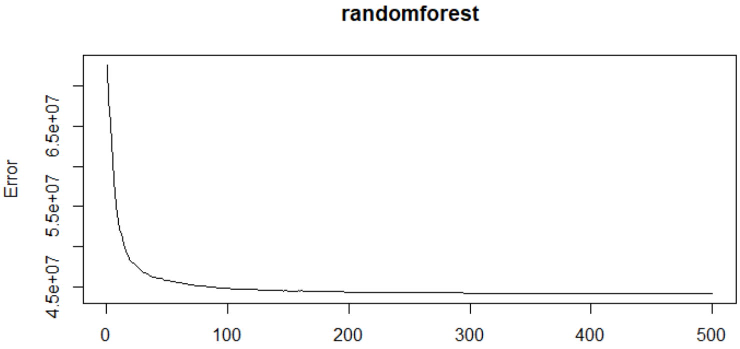


Figure 3: an example of the out of bag error for each number of trees in the random forest

The random forest models are run using the do.trace = 10 command in R, which provides the error of the model after every 10 trees that are constructed. The amount of trees is chosen by determining at which number of trees the error no longer decreases. At that point, constructing another tree will not result in a better model. The plot in figure 3 shows that the first 100 trees drastically decrease the out-of-bag (OOB) error of the forest. After roughly 300 trees, the OOB error stops decreasing. This means that, for the random forest in this example, using 300 trees would suffice.

Propensity score matching

For propensity score matching, a binary variable is needed that indicates whether an individual was 'treated' or not. In this case, there is no such variable, as all clients in the sample received district nursing care. Because of this, the effect of receiving district nursing care cannot be determined. This is why for this method, the effect of receiving treatment from a specific provider will be determined. For this method, a new variable is created that indicates whether a person was treated by one specific provider. The provider that is chosen as an example is provider X, a provider of district nursing care with a large client population. Because one provider has to be chosen as an example for this method, it is not possible to calculate the performance measures in the same way as for the other methods: for the other methods, a training and testing set can be made. This does not work for propensity score matching. This is why propensity score matching will only be present in the second part of the results section, where one provider is run through all the models as an example.

Logistic regression is used to estimate the propensity score for every client in the dataset. In this case, the propensity score can be interpreted as the chance that someone is treated by provider X on their characteristics.

For the matching algorithm, nearest neighbour matching was chosen, as this is the most straightforward matching estimator. To check whether the common support condition is fulfilled, and the balance of the matching is adequate, the standardised mean difference between the treated group and the control group is calculated for each covariate. There is no clear consensus on what threshold should be used for the standardised mean difference. Some researchers suggest that a standardised mean difference larger than 0.1 indicates a meaningful imbalance, so this is the threshold that will be used in this thesis [34]. Matching is deemed to be balanced when the standardised mean difference of all covariates is below 0.1. The same is true for the common support condition, as matching without a region of common support would have resulted in unbalanced matching [33]. Matching was performed without replacement, as this resulted in balanced matches. Matching with replacement is useful when the region of common support is very small, which would result in unbalanced matches [33].

As recommended by Caliendo, untreated subjects were oversampled, which means each treated subject was matched to multiple untreated subjects [33]. The degree of oversampling was determined by choosing the highest degree of oversampling that results in balanced matches. This ensures the lowest variance, while still maintaining a prediction of good quality.

Case study of provider X

To provide an example of how the predictions of the different models can be compared, one provider with a large number of clients was chosen. All models were trained again, this time excluding all clients from that provider. The models were then used to make a prediction for all clients of the chosen provider. This simulates the use of the benchmark model to benchmark a new provider.

Dimensions

The first dimension that the models will be assessed on is the variable selection. Firstly, the variable importance will be reported for each of the different models, as they all determine the importance of variables in their own way. A top ten of the most important variables according to each method will be presented. For the linear regression, variable importance will be assessed by using the coefficient and p value of a variable. A large coefficient (either positive or negative) and a small p value indicate an important variable. For the elastic net regression, coefficients are used to determine variable importance. For the random forest, a variable importance plot is made, which ranks the variables from most to least important. Lastly, for the regression tree, the level at which a variable is present in the tree is reported, as more important variables will be found higher up in the tree [28]. Next to presenting the top ten for the different methods, the importance of other variables will be discussed. For practicality, variable importance will be assessed using the version of the models that is applied to the entire dataset.

After this, the effects of the differences in the variable selection between the existing benchmark model and the new benchmark models mentioned above will be reported. This will be done by providing the coefficients and p values of the variables that differ between the old and new models. For the new models, these coefficients and p values will be extracted from the linear regression model on the whole dataset. This is done because for some of the other methods, such as the regression trees or propensity score matching, extracting coefficients and p values is not possible. P values indicate whether a variable has a significant effect on the outcome of interest. When the p value of a variable is lower than the chosen threshold, the null hypothesis can be rejected. In this case, the null hypothesis would be that the variable has no effect on the outcome of interest. The most commonly used threshold for the p value is $p < 0.05$, which is why this is the threshold that is used in this thesis.

The second dimension is the fit of the model. This will be assessed by calculating the MAE, RMSE and R^2, which were discussed in the theoretical framework for each model.

The third and final dimension the models will be assessed on is their usability. This dimension encompasses the degree to which the different models are explainable to both the people at Menzis responsible for healthcare purchasing, and the providers of district nursing. This will be measured by giving a presentation to healthcare purchasers at Menzis. The slides of the presentation can be found in appendix 6. During the presentation, the purchasers were first asked to think about what information a district nursing benchmark should be able to provide to best assist them during talks with providers of district nursing. After this, the methods that are applied in this thesis are explained to the purchasers. Elastic net regression and random forests were not explained to the purchasers, as these methods use the same mechanisms to predict as linear regression and regression trees respectively. For each remaining method, the purchasers were provided with an explanation of how each method comes to a prediction, and how each method can be interpreted. Lastly, the purchasers were asked to make a ranking of the different methods, and to place it in the chat of the online session. Then, they were asked to explain the reasoning behind their ranking. When every purchaser had given and explained their ranking, a discussion on their reasoning followed. After this discussion, the purchasers had the chance to alter their ranking before making it final. The presentation and the discussion that followed were recorded to aid in processing the results.

Results

Variable importance

This section is based on the output of the linear and elastic net regression, the variable importance plot of the random forest and the plot of the regression tree. These can be found in full in appendices 2, 3, 4 and 5 respectively.

Linear and elastic net regression

The top ten most important variables according to the linear regression consists of two age group variables, and eight health status variables. Age groups 20 (95-100 years of age) and 21 (100-105 years of age) are found in the top ten, the former being the fourth most important variable and the latter the single most important. Three diagnoses-based cost groups are present in the top ten, namely cost groups 12, 8 and 9, in order from highest to lowest coefficient. These variables are the second, eighth and tenth most important variables respectively. The variable that indicates whether a client received district nursing care in the previous year was found to be the third most important variable. Pharmacy-based cost groups 18 (brain and spinal cord diseases, other) and 15 (diabetes type 1a, with hypertension) occupy spots five and nine respectively. Lastly, assistive tool cost groups 2 (therapeutic elastic stockings) and 5 (tools for urine collection) were identified as the sixth and seventh most important variables.

Seven out of the ten variables that were mentioned above are also present in the top ten according to the elastic net regression. The three variables that are present in the top ten of the linear regression, but not in that of the elastic net regression are the pharmacy-based cost group regarding brain and spinal cord diseases, and diagnoses-based cost groups 8 and 9. While the coefficients of the first two of these were still present in the elastic net regression, the coefficient for diagnoses-based cost group 9 was reduced to 0. Instead of the three variables mentioned before, the top ten according to the elastic net regression was completed with the indicator whether a client died in the first six months of 2023, assistive tool cost group 0 (indicates whether someone used no assistive tools), and age group 19 (90-95 years of age). These first two of these variables are statistically significant in the linear regression model, but outside the top ten, while age group 19 is not significant. This means that the top ten most important variables according to the elastic net regression consists of seven health status variables, and three age group variables. Out of all variables mentioned in the two paragraphs above, assistive tool cost group 0 is the only one with a negative coefficient. This means that clients that fall in this cost group have lower district nursing costs than clients that do not. All other variables have a positive coefficient.

Both methods found the gender variable to be statistically significant, with women having higher costs than men. For the income variable, linear regression found the three lowest levels to be statistically significant. Elastic net regression reduced the coefficients of all levels to zero, except for the €18,000-€26,000 and €75,000-€100,000 groups. The coefficient of the latter is negative, while the coefficients for every other statistically significant level of income are positive. None of the levels of the education level variable were found to be statistically significant by the linear regression. Elastic net regression provides a similar result, as all coefficients for education level are reduced to zero.

Out of the variables regarding family composition, one was found to be statistically significant by the linear regression. This concerns the variable ‘number of people above the age of 18 in household’. This is also the only variable regarding family composition of which the coefficient was not reduced to zero by the elastic net regression. Both methods predict clients with more people above the age of 18 in their household to have lower district nursing costs.

The linear regression model found 6 out of 15 diagnoses-based cost group variables to be statistically significant. The coefficients for these variables generally rise as the cost group number goes up: cost group 4 has a higher coefficient than cost group 2. This is in line with the intent of the diagnoses-based cost group model, as diseases that have higher costs associated with them are placed in cost groups with a higher number [41]. The elastic net regression reduced all but two diagnoses-based cost group variables to zero. Again, the cost group with the higher number also has the higher coefficient.

Out of the 35 pharmacy-based cost group variables, 19 were deemed statistically significant by the linear regression. Among these are all four types of diabetes, heart disease, depression and Parkinson's disease. The elastic net regression included 6 out of 35 cost groups. All of these are also statistically significant in the linear regression model, except for cost group 0. This variable indicates no usage of medication associated with chronic disease. Elastic net regression predicts lower costs for these clients, while the variable is not statistically significant in the linear regression model.

Out of the 11 assistive tool cost groups, six were found to be statistically significant by the linear regression model. Among these are oxygen devices and ostomy provisions, as well as the ones mentioned before. Elastic net regression reduced the coefficients of all assistive tool based cost groups to zero, except for three of them. These are ones that are also statistically significant in the linear regression model. Both methods find the indicator for no use of assistive tools to be statistically significant, with a negative coefficient. A client that has not used assistive tools in the previous year, will be predicted to have lower district nursing costs in the current year.

Random forest and regression tree

Whereas the top ten most important variables according to the linear and elastic net regression consists mostly of health status variables, the random forest finds demographic variables to be more important. Out of the ten most important variables, seven are demographic variables. The top ten is completed with three health status variables. The seven demographic variables in the top ten are the age group, income, client type, education level, number of people above the age of 18 in the household, gender, and lastly the indicator for multiple earners in the household. The three health status variables are the indicator for district nursing use in 2021, the indicator for no use of assistive tools, and lastly the therapeutic elastic stockings cost group. The importance of the different levels of categorical variables cannot be determined here, as these variables were not transformed into dummies when applying the random forest and regression tree.

The 20 other variables that can be found on the variable importance plot are all health status variables. They mainly consist of diagnoses-based and pharmacy-based cost group variables, as well as one assistive tool cost group, and the variable that indicates whether a client died in the first six months of 2023. The direction of the effect of a variable cannot be determined for this method, as the specific splits that were performed by the random forest cannot be viewed.

The optimal regression tree that was found for the entire dataset makes use of 12 variables. In contrast to the random forest, the tree replaced three demographic variables (client type, gender and the indicator for multiple earners) with health status variables. This means that the regression tree contains five health status variables that are not present in the top ten according to the random forest. Two of those are not present on the variable importance plot of the random forest altogether. Both of these are diagnoses-based cost group variables. One contains inflammatory bowel disease and diseases with similar health care costs. The other contains lung disease, and other diseases that have similar health care costs associated with them.

Overview

In table 3, an overview of the variable importance according to the different methods is presented. Colour coding is used to display the variable importance. For the linear regression, a variable is dark green when the model placed it among the ten most important variables. A variable is light green when the variable was not among the ten most important, but it was statistically significant ($p < 0.05$). Variables that were not statistically significant are coloured red. For the elastic net regression, a variable is dark green when it was placed among the 10 most important variables. A variable is coloured light green when the variable was not present in the top ten, but its coefficient was not reduced to zero. The colour red corresponds to a variable of which the coefficient was reduced to zero.

For the random forest, dark green variables are those that the method found to be among the 10 most important variables. Light green variables were present on the variable importance plot, but outside of the top ten. Red variables were not present on the variable importance plot altogether. Lastly, for the regression tree: variables that are coloured light green were present in the tree, while variables that are coloured red were not.

By looking at table 3, the most important variables can be identified. The age group, district nursing in 2021 and use of therapeutic elastic stockings variables are included in the ten most important variables for each method, and are used by the regression tree. After this, the indicator for no use of assistive tools, the cost group for diabetes type 1 with hypertension, the cost group for urine collection devices, and the income variable perform best. These variables are placed in the top ten by at least one method, while being present or statistically significant in the other methods. Lastly, there are some variables that are not used by the regression tree, but are present or statistically significant in all other methods. This concerns gender, the deceased indicator, the client type, and the cost group for heart disease.

There are some variables that one of the methods identified to be among the top ten most important variables, while other methods did not find it to be useful. This is the case for multiple earners indicator and the cost group for HIV/AIDS. This occurs in a lesser degree for the cost group for oxygen devices, which is used by the regression tree in addition to the former statement. One other thing that stands out is that the random forest shows a preference for demographic variables, while linear and elastic net regression prioritise health status variables more often.

	Linear regression	**Elastic net regression**	**Random forest**	**Regression tree**
Demographic variables				
Gender	🟢	🟢	🟩	🟥
Age group	🟩	🟩	🟩	🟢
Income	🟢	🟢	🟩	🟢
Education level	🟥	🟥	🟩	🟢
Number of people above 18 in household	🟢	🟢	🟩	🟢
Number of people below 18 in household	🟥	🟥	🟥	🟥

Multiple earners indicator	🟥	🟥	🟩	🟥
Client type	▨	▨	🟩	🟥
Health status variables				
Deceased indicator	▨	🟩	▨	🟥
District nursing in 2021 indicator	🟩	🟩	🟩	▨
No use of medication	🟥	▨	▨	▨
No use of assistive tools	▨	🟩	🟩	▨
Cost group heart disease	▨	▨	▨	🟥
Cost group diabetes type 1 with hypertension	🟩	🟩	▨	▨
Cost group antidepressants	▨	🟥	▨	🟥
Cost group Parkinson’s	▨	🟥	🟥	🟥
Cost group ostomy provisions	▨	▨	🟥	🟥
Cost group therapeutic elastic stockings	🟩	🟩	🟩	▨
Cost group urine collection devices	🟩	🟩	▨	▨
Cost group oxygen devices	🟩	🟥	🟥	▨
Cost group inflammatory bowel disease	🟩	▨	🟥	▨
Cost group HIV/AIDS	🟩	🟥	🟥	🟥
Cost group lung disease	🟩	🟩	🟥	▨

Table 3: overview of the variable importance according to the different methods

Differences between the Menzis model and the new models

In table 5, the differences between the existing model and the new models regarding variable selection are presented. In the rightmost column, the coefficients and p values for the differing variables are provided. These values are taken from the linear regression model that was run on the full dataset. To determine whether effects are statistically significant, a threshold of 0.05 was chosen for the p value.

It was found that having more people above the age of 18 in the household results in lower district nursing costs. It can be concluded that having more people above the age of 18 in the household reduces district nursing costs (p<0.001). Having more people under the age of 18 in the household results in an increase of district nursing costs, ceteris paribus. However there is no indication that this effect is statistically significant (p=0.37). This suggests that splitting up the variable according to the age of the household members for the new models was an appropriate choice.

The sum of pharmacy-based cost groups variable that was added to the existing benchmark model had a positive correlation with district nursing costs: when a client uses medication from more than 1 pharmacy-based cost group, ceteris paribus, their district nursing costs increase. However, the effect of this variable on district nursing costs is not statistically significant (p=0.37). This suggests adding this variable to the benchmark model is not appropriate.

For the socioeconomic class variable which was included in the existing model, the software chose social class D as the reference level. This means that the effect of being in social class D is equal to 0. Clients with a higher social class are expected to have higher district nursing costs compared to class D, except the clients with social class C: clients within this social class have lower district nursing costs. The p values for each of the levels of this variable lie between 0.24 and 0.94, which means that none of the levels have a statistically significant effect on district nursing costs. This suggests that not including this variable in the new models was an appropriate choice.

The interaction variable regarding gender and age that is present in the existing model has a small coefficient of 1.9E-9. This is because the value of this interaction variable can be very large, as the age of a client is raised to the power of 6. This results in values as high as 1.00E+12 for this variable. The effect this variable has on district nursing costs is statistically significant (p<0.001). As interaction variables were outside of the scope of this thesis, conclusions cannot be drawn about whether including this interaction term would be appropriate. However, its high statistical significance suggests that it might be appropriate to include.

For the variable that indicates whether more than one person in the household has a job that was added to the new models, the level ‘no partner, and/or not classified’ was chosen as the reference variable. Both having and not having multiple earners in a household resulted in an increase in district nursing costs. At p values of 0.15 and 0.20 respectively, this effect is not statistically significant. Thus, including this variable in the new models does not seem to be appropriate.

Receiving district nursing care in the year prior, ceteris paribus, resulted in an increase of district nursing costs for the new year by €3990. This effect is statistically significant (p<0.001). This looks to be one of the most important variables for predicting district nursing costs. Thus, it looks to be an appropriate addition to the new models.

Menzis model	**New models**	**Coefficient and p value from linear regression**
Members in household	Members of household below and above the age of 18 separately	Under 18: 130, P = 0.37[1] Above 18: -269 P < 0.001[1]
Sum of pharmacy-based cost groups added	No sum of pharmacy-based cost groups added	1611, P = 0.37[2]
Socioeconomic class added	No socioeconomic class added	Class A: 404, P = 0.39[2] Class B1: 470, P = 0.24[2] Class B2: 361, P = 0.29[2] Class C: -13, P = 0.94[2]
Gender * age^6 added	No interaction variable added	1.9E-9 P < 0.001[2]
No indicator for multiple people with a job in household added	Indicator for multiple people with a job in household added	Yes: 251, P = 0.15[1] No: 194, P = 0.20[1]
No indicator for district nursing use in past year added	Indicator for district nursing use in past year added	3990, P < 0.001[1]

Table 3: the differences between the existing benchmark model and the new model. Footnotes indicate whether the coefficient stems from the existing or new model.

Fit of the models

In table 6, the results regarding the fit of the different models are presented. As each method was applied to the different subgroups which were made according to the number of months that district nursing was used by a client, results are presented for each subgroup. For each of these subgroups, the new model that performs best on the different performance measures is bolded. In the first column, some information about the different parameters that have to be chosen is provided. For the elastic net regression, the values for α and λ are provided. For each regression tree, the number of leaf nodes in the tree is shown. Lastly, for the random forests, the number of trees in the forest is reported.

When applying the models to the entire dataset, the existing model shows the lowest MAE and RMSE, but also the lowest R^2 out of all the models. This is likely explained by the fact that the existing model was trained without the 4% of clients with the highest district nursing costs. This means that the existing model cannot be properly compared to the new models by using the MAE and RMSE. The R^2 does not suffer from this. The linear regression had the highest R^2 for this category. All new models had a higher R^2 than the existing model for this category. All new models performed similarly to each other.

For each model, the values for R^2 are the highest for the entire dataset. For the subgroups, the values for R^2 are the lowest for the group of 1 to 3 months of care received. The values for R^2 rise when the number of months of care received increases. This suggests that the models perform best when applied to the entire dataset, and worse when applied to subgroups. The fact that the values for R^2 rise for the subgroups that received district nursing care for a longer timespan suggests that costs for clients that received short term care are more unpredictable than for clients that received chronic care. The MAE and RMSE values are the lowest for the 1 to 3 months subgroup, and the highest for the 10

[1] Coefficient and p value from new model
[2] Coefficient and p value from Menzis model

to 12 months subgroup. This is because district nursing costs tend to increase when a client receives care for longer periods of time.

For both the entire dataset and all subgroups, the existing model shows the lowest values for both the MAE and RMSE. The difference between the RMSE values of the new models and the existing model is larger than the difference between the MAE values. This provides information about the distribution of errors: large errors occur more often when using the new models. This can be deduced from the fact that the RMSE gives a relatively high weight to large errors.

The R^2 values for the 1 to 3 months subgroup are the lowest for each of the models. This means that for this subgroup, the smallest amount of variance is explained by the models. The existing model performs best for this subgroup, with both the lowest MAE and RMSE, and the highest R^2. The new models perform similarly for this subgroup, with the exception of the regression tree: this model has a lower R^2. For this subgroup, the MAE values for the existing model, linear regression and elastic net regression are equal to each other, while the RMSE values differ: the existing model has a lower RMSE value than the linear regression and elastic net regression. This indicates that, for the latter two methods, there are more predictions with higher errors than for the existing model.

For the 4 to 6 months subgroup, the existing model again slightly outperforms the new models, with the lowest MAE and RMSE values, and the highest R^2. Again, the new models perform similarly, with the exception of the regression tree, which has higher error rates and a lower R^2 than the other models.

When looking at the 7 to 9 months subgroup, the existing model performs the best on all performance metrics. Here, the difference in R^2 is larger: the R^2 for the existing model is more than twice as large as that of the linear regression and elastic net regression. For this subgroup, the random forest has both the highest MAE, RMSE and R^2 values out of the new models.

For the 10 to 12 months subgroup, the linear regression performs the best on all performance metrics out of the new models. Its R^2 is also higher than that of the existing model.

Method **Entire dataset**	**MAE**	**RMSE**	$\boldsymbol{R^2}$
Existing model	3499	4729	0.157
Linear regression	4415	**6754**	**0.209**
Elastic net regression (α = 0.8, λ = 250)	4454	6811	0.199
Regression tree (n end nodes = 25)	4357	6840	0.191
Random forest (ntree = 500)	**4345**	6781	0.206
Propensity score matching	Not applicable	Not applicable	Not applicable
1-3 months of district nursing care			
Existing model	735	1165	0.018
Linear regression	**735**	**1241**	**0.014**

Elastic net regression (α = 0.1, λ = 91)	**735**	1249	0.013
Regression tree (n end nodes = 28)	748	1303	0.001
Random forest (ntree = 100)	754	1261	0.012
Propensity score matching	Not applicable	Not applicable	Not applicable
4-6 months of district nursing care			
Existing model	1923	2676	0.049
Linear regression	2026	2930	**0.047**
Elastic net regression (α = 0.2, λ = 92.3)	**2018**	**2927**	**0.047**
Regression tree (n end nodes = 93)	2160	3129	0.017
Random forest (ntree = 500)	2052	2948	0.042
Propensity score matching	Not applicable	Not applicable	Not applicable
7-9 months of district nursing care			
Existing model	3428	4446	0.095
Linear regression	3773	5230	0.046
Elastic net regression (α = 0.1, λ = 873)	**3749**	**5156**	0.043
Regression tree (n end nodes = 24)	3879	5507	0.020
Random forest (ntree = 1000)	3969	6046	**0.058**
Propensity score matching	Not applicable	Not applicable	Not applicable
10-12 months of district nursing care			
Existing model	4546	5485	0.094
Linear regression	**6320**	**8497**	**0.105**
Elastic net regression (α = 0.5, λ = 846)	6423	8644	0.081
Regression tree (n end nodes = 19)	6481	8734	0.061

Random forest (ntree = 100)	6435	8646	0.084
Propensity score matching	Not applicable	Not applicable	Not applicable

Table 4: the fit of the different models, using three performance measures

Case study of provider X

In table 7, the results of the case study of provider X are presented. The results can be interpreted as follows: the -167 for the existing model in the entire dataset group means that, according to the existing model, provider X reported costs that were on average €167 lower per client than expected based on the characteristics of their patient population. In the rightmost column, a 95% confidence interval around the estimate is provided.

For the entire dataset, all models report that provider X has, on average, lower costs per client than expected. The estimates range from €-167 for the existing model, to €-357 for the random forest. For each estimate, the full confidence interval is lower than 0. This means that it can be concluded that provider X has lower costs than expected when looking at their entire patient population.

When looking at the 1 to 3 months subgroup, again all models show lower average costs per client than expected. Here, the variance between the estimates is smallest, ranging from €-154 for the existing model, to €-184 for the regression tree. Just like for the entire dataset, all confidence intervals are entirely below 0, which means that the difference between the actual and expected costs is statistically significant.

For the 4 to 6 months subgroup, just like before, all models show lower average costs per client than expected. The estimates range from €-132 for the linear regression, to €-188 for the elastic net regression. However, for this subgroup, all confidence intervals span both negative and positive values. This means that, for this subgroup, it cannot be concluded that the difference between the actual and expected average costs per client is statistically significant.

For the 7 to 9 months subgroup, every model reports lower average costs per client than expected. The variance in the estimates is the biggest for this subgroup: they range from €-366 for the existing model, to €-673 for the elastic net regression. The confidence intervals for the linear regression, elastic net regression and random forest lie below 0. This means that if you were to base your decision on one of these models, you would conclude that the average costs per client of provider X are significantly lower than expected. However, for the existing model, regression tree and propensity score matching, the confidence intervals also contain positive values. The conclusion based on one of these models would be that the average costs per client from provider X do not significantly differ from the expected costs.

The 10 to 12 months subgroup is the only subgroup were the models estimate that provider X declares higher average costs per client than expected. The estimates range from €142 for propensity score matching, to €345 for the regression tree. All confidence intervals for this subgroup span both positive and negative values. This means that the costs that provider X declared from this subgroup do not significantly differ from the expectation.

For all subgroups, the estimate of the existing model is either the one closest, or one of the closest to 0. On the other hand, there is no clear trend in which model provides the most extreme estimate.

Method	$\frac{(\Sigma_{i=1}^{N} \, actual\ costs_i - \Sigma_{i=1}^{N} \, predicted\ costs_{i)}}{N}$, **where N is the number of clients of provider X**	**95% confidence interval (z = 1.96)**
Entire dataset		
Existing model	-167	[-305, -29]
Linear regression	-290	[-485, -95]
Elastic net regression	-320	[-519, -121]
Regression tree	-322	[-520, -122]
Random forest	-357	[-553, -161]
Propensity score matching	-298	[-554, -42]
1-3 months of district nursing		
Existing model	-154	[-202, -106]
Linear regression	-173	[-221, -124]
Elastic net regression	-167	[-261, -119]
Regression tree	-184	[-239, -129]
Random forest	- 175	[-224, -126]
Propensity score matching	-163	[-224, -101]
4-6 months of district nursing		
Existing model	-139	[-349, 71]
Linear regression	-132	[-342, 78]
Elastic net regression	-188	[-397, 23]
Regression tree	-159	[-373, 57]
Random forest	-175	[-381, 33]
Propensity score matching	-146	[-398, 105]
7-9 months of district nursing		
Existing model	-366	[-757, 25]
Linear regression	-493	[-943, -43]
Elastic net regression	-673	[-1110, -232]
Regression tree	-419	[-890, 52]
Random forest	-555	[-996, -114]
Propensity score matching	-630	[-1288, 28.6]
10-12 months of district nursing		
Existing model	210	[-73, 491]
Linear regression	297	[-158, 750]
Elastic net regression	242	[-223, 705]
Regression tree	345	[-122, 810]
Random forest	190	[-273, 651]
Propensity score matching	142	[-400, 685]

Table 5: results of the case study of provider X

Usability

The presentation to assess the usability of the different models was presented to four people that are involved in the purchasing process regarding district nursing care. Of the attendants, one is a financial expert at Menzis, one a health care expert, and the last two are (senior) health care purchasers. As the team that is responsible for health care purchasing regarding district nursing is small, it can be concluded that this sample is representative of health care purchasers involved in district nursing at Menzis. After the discussion during the presentation, the health care purchasers unanimously came to the following ranking: matching was deemed the most suitable method, and linear regression the least suitable. The main argument for choosing matching was that there is no model involved in predicting the district nursing costs of clients: users of district nursing are matched to other real clients. The purchasers believe using real clients as comparators results in the most pure comparison. They argue that district nursing providers will be much more receptive to feedback based on a benchmark when matching is used, than when predictive models like linear regression are used. The purchasers note that there should be an agreement in place on what degree of oversampling of untreated subjects is necessary when using matching methods. They believe that when a predicted variable originates from only a few matched subjects, this prediction is not robust. They want a client to be matched to as many clients as possible to generate the most robust prediction.

The purchasers believe that variables which provide information about the providers should be included in the benchmark model. This information could include things such as the number of employees, and the average education level of these employees. This way, they can provide district nursing providers with more specific feedback on their performance.

One purchaser mentioned that providers are much more open to feedback based on a benchmark that makes comparisons within their organisation. An example of this would be to benchmark the different teams within their organisation against each other. When benchmarking like this, providers could be told that, to improve their overall performance, they should pay attention to the modus operandi of one specific team within their organisation. The purchaser believes that providers perceive this type of feedback to be more useful than a comparison to other providers.

The purchasers believe that making use of 'profiles' of characteristics would be helpful while benchmarking. They believe providers can be given more specific feedback in this way. An example of the feedback that could be given to a provider when using profiles is as follows: 'we see that you report higher costs than expected for older, male patients without a partner'. This type of feedback corresponds the most with the use of tree-based predictions, such as the regression tree and random forest. This is because when using profiles, the interaction between variables is important: in the example before, it could be the case that living without a partner has a higher effect for older male patients than for younger female patients. Tree-based models are easy to use in a case like this, as they find the optimal splits that capture these interactions by themselves. In other models, such as linear or elastic net regression, these interactions would have to be specified by the person performing the analysis.

Discussion

Main findings

According to the results regarding variable importance, both demographic and health status variables are among the most important when predicting district nursing costs. Out of the demographic variables, age, gender, income, client type and income were found to be most important by the different methods. The indicators for use of district nursing care in the past year and death in the first half of the next year were found to be the most important health status variables. Various cost group variables were also found to be important, such as the cost groups for therapeutic elastic stockings and diabetes type 1 with hypertension, among others. When assessing the variable importance, the results of the regression tree were not taken into account. This is because the regression tree is prone to overfitting, which means its results regarding variable importance are not generalisable [29].

The results concerning variable selection suggest that a combination of the variable selection of the existing model and the new models would work best. For example, the results suggest that splitting up the number of people in a household based on age was appropriate, as the number of people above the age of 18 significantly affects costs, while the number of people under the age of 18 does not. On the other hand, the inclusion of a variable that indicates whether more than one person in the household has a job does not seem appropriate, as the effect of this variable was not statistically significant. This means that it cannot be concluded whether this variable influences district nursing costs. This is an aspect where the variable selection of the existing model is better than that for the new models.

For the fit of the model, it was found that the existing model shows the lowest values for the MAE and RMSE across all subgroups. This can most likely be explained by the fact that outliers were removed when applying the existing model: the 4% of clients with the highest costs in each age category were excluded. Generally, outliers should only be removed from the analysis if they are the result of errors in the data. There was no indication that the outliers in the Menzis case were caused by errors in the data, which is why they were not removed for the new models presented in this thesis. The existing model showed the lowest value for R^2 out of all the models when applied to the entire dataset. All new models produced a similar R^2 value, which suggests they have similar predictive value. Chicco et al. reported that R^2 is the most informative performance measure when evaluating regression analysis, as R^2 does not suffer from the same interpretability problems as the MAE and RMSE [43]. MAE AND RMSE cannot be directly compared across different models, as they are dependent on the scale of the outcome variable. Because of this reason, the existing model cannot be compared to the new models using these performance measures, as the existing model was trained and tested using different data. The existing model shows the highest or second highest R^2 for each of the other subgroups. This is likely explained by the fact that the existing model was not only trained, but also tested with data where the outliers were removed. It is unclear whether testing the existing model while including these outliers would yield different results.

When looking at the fit of the model dimension, it can be concluded that the existing model performs the best for all subgroups of the data where clients were categorised according to the number of months of district nursing care they received. For the analysis where the entire dataset was used, the new models performed best, with the exception of the regression trees. The fact that the regression trees performed worse can be explained by the fact that singular regression trees are prone to overfitting to the training data [29]. Random forests also makes tree-based predictions, but do not suffer from this problem due to the fact that a random forest consists of a large number of regression trees. This is also reflected in the results, as the random forest always outperformed the regression tree. The fact that the values R^2 are the lowest for the 1 to 3 months subgroup, and rise when the number of months of care increases, suggests that short term care is more unpredictable than long term care. Across all subgroup analyses, the new models performed similarly, with the exception of the regression trees. This suggests that, based on the fit dimension, each of the new models, bar the regression trees, is roughly equally suitable for Menzis. Due to the way in which propensity score matching was applied, it was not possible to assess the fit of this method. This will be further discussed in the weaknesses section below.

All models have the highest value for R^2 when they are applied to the entire dataset, and the R^2 values drop drastically for the other subgroups. This suggests that the models perform quite well at distinguishing clients that receive short term care from clients that receive long term care. A possible explanation for this is that the variance within subgroups is higher than the variance between subgroups.

The results regarding the opinion of the users, and usability of the models show a clear preference towards propensity score matching. The main argument for this is that the purchasers believe providers are much more likely to respond positively to feedback, when the feedback originates from a benchmark that matches clients to similar, real clients. They prefer this over having a model which predicts district nursing costs. Tree-based methods would be the second preference of the purchasers, as these methods are the best at accommodating for the use of client profiles. The purchasers believe that using these client profiles will result in providers being more receptive to feedback. What is particularly interesting is the fact that linear regression was determined to be the least suitable method by the health care purchasers, when this is the method that Menzis most commonly uses in its analyses.

Based on these results, it can be concluded that either propensity score matching or the random forest is the most suitable method for the district nursing benchmark. The caveat to this is that propensity score matching could not be evaluated on the fit dimension. Propensity score matching is expected to have the highest impact in practice, due to the fact that providers of district nursing are expected to be more receptive to this method. The random performs similarly to linear and elastic net regression in terms of fit, but performs better on the usability dimension. The opinion of the end user is a very important determinant of the successful implementation of a technology. As described in the Technology Acceptance Model, some important determinants of successful implementation of a technology are the perceived usefulness and perceived ease of use [44]. This means that if the intended users of the technology do not perceive it as a helpful and easy addition to their work, implementation of the technology will likely fail. Random forests are the second most suitable option, because of a combination of the results regarding fit and usability. The fact that tree-based methods ended in second place of the ranking, makes the random forest the next candidate. This is reinforced by the fact that the random forest performed similarly to the other new models regarding the fit.

Strengths

The main strength of this thesis is that a wide range of machine learning methods were applied to the Menzis case. These methods were compared on multiple dimensions: their variable selection, their fit and their usability. This allows for a multifaceted advice to be given. The fact that the usability dimension was included makes the final advice significantly more useful to Menzis. Because the advice that is given to Menzis also takes the opinions and desires of the users into account, the implementation of this advice will likely go more smoothly than when this was not taken into account.

Next to the fact that various methods were applied, these methods were also applied in depth. For example, for elastic net regression, cross-validation was used to determine the optimal values for α and λ. For the random forest, the optimal number of trees was found by observing the out-of-bag error. Lastly, when applying propensity score matching, the balance of the matches was assessed through observing the standardised mean difference of each variable.

Weaknesses

One of the strengths of this thesis also brings a weakness: all methods applied in this thesis were compared to the existing benchmark model, but due to the differences in variable selection and data manipulation, the models are not completely comparable. The main differences between the existing model and the new models is the fact that the existing model removed outliers and all deceased clients, while the new models did not. A way to examine the effect of the differences between the existing and new models would be to run all models with the same variable selection and data manipulation as the existing model. After this, one of the differences that was presented in this thesis can be applied. This way, the entire difference in the results would be explained by the difference that was applied. This process could then be repeated for each difference regarding variable selection and data manipulation. By using this method, the effect of each difference regarding variable selection and data manipulation could be assessed. This was not done for this thesis due to time constraints.

A second weakness concerns the fact that some variables that are included in the benchmark are based on estimations: Menzis acquired these data from a third party. This concerns the following variables: income, education level, personality of the client, and the number of people in the household. Ideally, only factual variables should be included in the benchmark models. This means that replacements will have to be found for the estimated variables. An example of replacing an estimated variable by a factual one would be to replace the estimated number of people in a household by the amount of policyholders at an address. The latter is information that Menzis collects for every policyholder. However, including this variable would not fully solve the problem: different people in a household might not share their health insurance. This means that this variable would not accurately reflect the number of people in a household. This would not necessarily be a problem if the errors in this variable are randomly distributed, but this is probably not the case here. It is plausible that people with a higher education level pay more attention to which health insurance fits their needs, so the chance they end up on the same policy as their partner is lower than for people with a lower education level. This would mean that the amount of policyholders will more often wrongly reflect the amount of people in a household for people with higher education levels, than for people with lower education levels. Thus, the errors would not be randomly distributed, which would skew the predictions. When there is no factual variable readily available that would still correctly reflect the information of interest, these estimated variables are the best option.

A variable with information about the personality of clients was included in the benchmark models in this thesis. But should you correct for the personality of clients? As mentioned above, this variable is estimated, which is not ideal. This means that the variable was measured for a part of the population, and then extrapolated to the rest of the population based on their characteristics. Secondly, there is a difference between the care that a patient needs, and the care that the patient asks for. For example, a luxury oriented client might request more care than is really necessary. If you were to correct for the personality of clients, a provider with a large share of luxury oriented clients will have higher expected costs. This way, a provider is 'rewarded' with a good benchmark score for providing unnecessary care, while they should ideally be punished for this. The opposite can also occur, for example with the docile clients. A docile client might not request all the care they need, so the expected costs for those clients might be lower. A provider with a large share of docile clients could then be 'punished' for providing these clients the care they need. This means that the personality of a client should not be corrected for in the benchmark model. However, the variable can still be of use to Menzis. One possible use for the variable is to look at the personality mix of clients of one provider. An example of a piece of advice that Menzis could give to a provider based on that is that a provider should pay attention that they do not provide clients with more care than they need, because there are a lot of luxury oriented clients in their population.

When evaluating predictive models, it is common to use some kind of performance measure that takes into account the complexity of the model. Examples of such a measure include the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) [45,46]. These performance measures were not used in this thesis, as it is either impossible or impractical to calculate these for some of the methods presented in this thesis. For example, the application of the AIC and BIC to linear regression is quite straightforward: the values are determined by the number of explanatory variables that are included, as well as the fit of the model. However, using only the amount of explanatory variables does not work for a random forest. The complexity of a random forest is determined not only by the amount of explanatory variables, but also by the amount of splits that are attempted at each node and the amount of trees in the forest. Thus, the AIC and BIC cannot be calculated for this method. Even if this calculation was possible, it would not provide any additional information, as a random forest is inherently more complicated than a linear regression. Lastly, the argument of reducing the cost of gathering data does not hold up in this situation, as the benchmark models are built using data that Menzis gathers every year for multiple purposes.

A weak point regarding the application of elastic net regression is the fact that the ideal values for α varied strongly when using different train and test sets. While 10-fold cross-validation was used to find the best value for α each time elastic net regression was applied, it seems that a more robust form of cross-validation will have to be used. This way, there will be more certainty that the ideal value for α that is found is, in fact, the actual ideal value. The possible application of k-fold cross-validation to the train and test sets will be discussed in the future work section of this discussion.

Another weak point that can be identified in the application of elastic net regression is the fact that there is no indication of significance next to the coefficients. This is because this method does not produce p values. One could say that the point of the elastic net is to reduce the coefficients of unimportant variables to zero, and that all variables that remain in the elastic net model are statistically significant. If it is determined that some indication of significance is needed for the elastic net model, there are other types of significance tests that could be applied, such as the post-selection inference approach Lee et al. describe [47]. This was not done in this thesis due to time constraints.

A final caveat to the application of elastic net regression is the way the method handles categorical variables. For each level of a categorical variable, the model decides whether to shrink the coefficient or not. This means that for a variable like income, only one out of multiple levels ends up in the final model. This is also what happened in this thesis, as can be seen in appendix 3. One of the reasons for using elastic net regression is the fact that it reduces the coefficients of unimportant variables to zero. This lowers the costs that are associated with gathering data, as now the data for the unimportant variables does not have to be gathered anymore. However, when only one level of income is included in the model, the cost of data gathering is not lowered. All data regarding income will still have to be gathered, even if only one level is used in the final model. This problem can be solved by using a grouped regularization method [48]. This method groups the levels of categorical variables together. Then, the elastic net algorithm is only allowed to shrink the coefficient of one group as a whole. This way, it is likely that a categorical variable with a large number of levels, of which only one is important, will be excluded as a whole. This would lower the costs associated with gathering data. The grouped elastic net was not applied in this thesis, because reducing the amount of data that has to be collected is not relevant in the case of Menzis. As mentioned before, the data that is used in this thesis are used for various applications, which means that they are gathered every year.

A weak point regarding the application of propensity score matching is that the different impact variables have on district nursing costs was not considered. This means that some variables that are irrelevant might be included in the estimation of the propensity scores. The matching algorithm will try to find matches that are balanced for all variables, including the irrelevant ones. Thus, achieving sufficient balance for the irrelevant variables might come at the expense of a worse balance for variables that are actually relevant predictors of district nursing costs. As a result of this worsened balance, more bias is introduced into the predictions. One possible solution for this is to ensure beforehand that all variables that are used to estimate the propensity scores are significant predictors of district nursing costs. This could be done using a regularization method, such as lasso or elastic net regression. When all variables that are used to estimate the propensity scores are significant predictors of district nursing costs, the risk of reducing the balance of a variable that is way more important is lowered. Thus, the risk of introducing bias is lowered. A further solution to this problem is to adjust the threshold for the standardised mean difference for each variable. This is because a small amount of imbalance in important variables can result in a large difference in the district nursing costs. This way, the threshold for the standardised mean difference for variables that have a large effect on district nursing costs will be lowered, while the threshold for variables that have a small effect on district nursing costs will be raised. This results in less bias.

The inclusion of the usability dimension also carries limitations with it. One of these is the fact that the opinions of the attendees of the presentation can be influenced by the contents of the presentation and the way in which they were presented. This means that the presenter can have a large influence on the opinions of the audience. This effect was negated by the fact that both external supervisors were involved in optimising the presentation. Secondly, the interpretation of the audience can make for a large difference in opinion. However, there is no indication that the interpretation of the attendees differed strongly from each other, as they all independently came to the same ranking. Furthermore, the degree to which providers of district nursing are receptive to feedback that is based on the different methods turned out to be an important factor. However, this was only measured by consulting health care purchasers, which means it was indirectly measured. A better approach would have been to consult providers of district nursing as well.

Lastly, a weakness can be found in the comparison that is made between the models. Propensity score matching was not assessed on the fit dimension, which makes it impossible to draw a full comparison between this method and the other models. Propensity score matching was not present in the results of the fit dimension, as this method can only predict the costs for one provider at a time: this is because propensity score matching makes use of a ‘treatment’ variable, which in this case indicates whether a client was treated by the specific provider that is being benchmarked. To be able to include propensity score matching in the results section for the fit dimension, the analysis would have to be run separately for each of the 386 different providers of district nursing in the dataset.

Future work

The focal point of any further inquiry into the performance of the different models presented in this thesis should firstly be the differences in variable selection and data manipulation between the existing model and the new models. The difference in the fit dimension between the existing model and the new models is likely caused by differences in data manipulation. These differences will have to be further examined to determine which steps of data manipulation are the best. As mentioned above, the variable selection part of the results section suggests that the best possible variable selection includes parts from both the existing model and the new models. Thus, the variable selection and data manipulation of both the existing model and the new models should be thoroughly re-evaluated. This can be done by starting with steps of which it is certain that they are appropriate. After this, changes of which the suitability is uncertain should be applied incrementally to test whether they should be taken.

As mentioned in the results, the values for the MAE and RMSE rise when the models predict costs for clients who received district nursing care for a larger number of months. This is due to the fact that a client's district nursing costs increase when they receive care for a larger number of months. Because of this, it becomes difficult to gauge whether the models performed better or worse when comparing the subgroups using the MAE and RMSE. This problem could be solved by adding another performance measure that can give information about the proportion of the errors. Examples of such performance measures are the mean absolute percentage error (MAPE) and its symmetric variant (SMAPE) [43].

A strong improvement that could be made in a further evaluation of the models presented in this thesis is the use of k-fold cross-validation [49]. When using this method, the data are randomly split up into k mutually exclusive subsets, also known as the fold. These folds are of approximately equal size. The predictive model is trained and tested k times, once for each different fold. Each time, the model is trained using every fold bar one, which is the fold the model is tested on. The different performance measures, such as the ones used in this thesis, are calculated for each different fold. After this, the average of the performance measures across all folds is taken to determine the final value for each measure. Using this method is more robust than only splitting the data into train and test sets once, and results in lower variance of the performance estimates. It also helps reduce overfitting, as the model is exposed to different subsets of the data.

Another area of improvement for further evaluation lies in the application of propensity score matching. In this thesis, nearest neighbour matching was the only matching technique that was applied. However, there are multiple other matching techniques that each have their own advantages, such as complete matching and calliper matching [33]. It could be investigated which matching algorithm fits the best to the Menzis case.

In this thesis, the data were split up into subsets according to the number of months a client received district nursing care. This was done to account for the difference between short term and long term district nursing care. A better way of doing this would be to classify clients as short term or long term by looking at their treatment codes. These codes can more accurately indicate whether a client received short or long term district nursing care. However, using these treatment codes is not practical yet at this point in time, as there are a lot of missing values in those data. When these data would be complete, it would be a good idea to use the treatment codes instead of the number of months of district nursing to account for the difference between short term and long term care.

The Menzis purchasers had two main wishes for the district nursing benchmark. However, both of these wishes bring the risk of introducing bias when applied to their favoured method, which is matching. Firstly, they think the provider dataset that Menzis possesses should be integrated into the benchmark model, so they can give more specific feedback to providers of district nursing on their performance. This however carries the risk of making it harder for the matching algorithm to find balanced matches, as there would be more variables to match on. When the variables regarding the provider are not included, the matching algorithm has to find a client with similar characteristics that receives care from any different provider. When the variables are included, the matching algorithm has to find a client with similar characteristics, that receives care from a different provider with similar characteristics. This will likely result in matches that are less balanced, which results in more bias. Thus, Menzis should be cautious when including these variables in the benchmark. If they decide that the inclusion of these variables is necessary, only variables that are relevant for predicting health care costs should be added. As mentioned above, adding irrelevant variables to the estimation of the propensity scores results in more bias. Another problem with adding the variables concerning providers is that a lot of these variables can be influenced by the providers. As mentioned in the theoretical framework, these influenceable variables should not be blindly added to the benchmark. They should either be removed from the benchmark altogether, or they should be adjusted for prospectively.

Secondly, when applying matching, the purchasers believe the degree of oversampling of untreated subjects should be as high as possible. This way, they find the expected value that the method provides to be more robust, and more explicable to providers. The second wish of maximising the degree of oversampling of untreated subjects involves the trade-off between bias and variance that was discussed in the theoretical framework [33]. The fact that this trade-off is in place means that Menzis should carefully consider the amount of oversampling that they apply when using matching, as increasing the amount of oversampling increases bias.

The use of 95% confidence intervals during the case study provided the insight that the models cannot always tell whether the real costs that a provider made are meaningfully different from the model prediction. This is the case when the confidence interval contains 0. The confidence intervals that were reported in the case study are an indication of the amount of uncertainty in the estimates. Currently, Menzis does not make use of confidence intervals when interpreting the results of the benchmark model. This could result in a provider being told that they are making higher costs than expected, while in reality their costs are lower than could be expected. Using a confidence interval could prevent this problem from occurring. For scientific research, using a 95% confidence interval is common. However, as Menzis does not use their benchmark model for scientific research, the same level of confidence might not be necessary. The implementation of any confidence interval is already an improvement compared to the current situation. Menzis will have to decide on what level of confidence they find sufficient to make decisions based on the results of the benchmark models.

Conclusion

Propensity score matching and random forests seem to be the most suitable methods to use for the district nursing benchmark of Menzis. This is backed by the fact that propensity score matching is the clear favourite among the health care purchasers, who are the end users of the benchmark. Random forests combine the second best score among the purchasers with a performance that is similar to the other models presented in this thesis regarding fit. The biggest recommendation for future evaluation of the methods presented in this thesis is to re-evaluate the variable selection and data manipulation steps of building the models. After that, the models could be further evaluated by applying methods such as k-fold cross-validation and different matching algorithms for propensity score matching.

References

[1] Dale B. Managing Quality, 2003, p. 427–40.

[2] Wilson A, Nathan L. Understanding benchmarks. Home Healthc Nurse 2003;21:102–7.

[3] Massoud R, Askov K, Reinke J. A Modern Paradigm for Improving Healthcare Quality. Quality Assurance Project; 2001.

[4] Maire JL, Bronet V, Pillet M. A typology of “best practices” for a benchmarking process. Benchmarking 2005;12:45–60. https://doi.org/10.1108/14635770510582907/FULL/PDF.

[5] Barendregt M. Benchmarken en andere functies van ROM: back to basics. Tijdschr Psychiatr 2015;57:517–25.

[6] Nederlandse Zorgautoriteit. Benchmark Ziekenhuizen. 2012.

[7] Rijksoverheid. Nederlandse Zorgautoriteit (NZa) | Contact | Rijksoverheid.nl n.d. https://www.rijksoverheid.nl/contact/contactgids/nederlandse-zorgautoriteit-nza (accessed March 6, 2023).

[8] Nederlandse Zorgautoriteit. Contracten tussen zorgverzekeraars en zorgaanbieders 2015.

[9] Menzis. Benchmark wijkverpleging 2022.

[10] Menzis. Jaarrekening en kerncijfers 2021. https://www.menzis.nl/over-menzis/jaarverslagen-en-kengetallen (accessed March 9, 2023).

[11] Zorginstituut Nederland. Wijkverpleging (18 jaar en ouder) | Verzekerde zorg | Zorginstituut Nederland n.d. https://www.zorginstituutnederland.nl/Verzekerde+zorg/wijkverpleging-zvw (accessed March 20, 2023).

[12] Hoeck S, François G, Geerts J, Van Der Heyden J, Vandewoude M, Van Hal G. Health-care and home-care utilization among frail elderly persons in Belgium. Eur J Public Health 2012;22:671–7. https://doi.org/10.1093/EURPUB/CKR133.

[13] Linden M, Horgas AL, Gilberg R, Steinhagen-Thiessen E. Predicting Health Care Utilization in the Very Old. Http://DxDoiOrgEzproxy2UtwenteNl/101177/089826439700900101 1997;9:3–27. https://doi.org/10.1177/089826439700900101.

[14] Centraal Bureau voor de Statistiek. Gezondheid en zorggebruik; persoonskenmerken, 2014-2021 2023. https://www.cbs.nl/nl-nl/cijfers/detail/83005NED (accessed April 23, 2023).

[15] Duncan I, Ahmed T, Dove H, Maxwell TL. Medicare Cost at End of Life. Am J Hosp Palliat Care 2019;36:705. https://doi.org/10.1177/1049909119836204.

[16] Kempen GIJM. Thuiszorg voor ouderen, Een onderzoek naar de individuele determinanten van het gebruik van wijkverpleging en/of gezinsverzorging op verzorgend en huishoudelijk gebied 1990.

[17] De Meijer CAM, Koopmanschap MA, Koolman XHE, Van Doorslaer EKA. The role of disability in explaining long-term care utilization. Med Care 2009;47:1156–63. https://doi.org/10.1097/MLR.0B013E3181B69FA8.

[18] Wingen M, Otten F. Sociaaleconomische status en verschillende gezondheidsaspecten van ouderen. Tsg 2009;87:109–17. https://doi.org/10.1007/bf03082194.

[19] Andersen R, Newman JF. Societal and Individual Determinants of Medical Care Utilization in the United States. Milbank Q 2005;83:Online-only. https://doi.org/10.1111/J.1468-

0009.2005.00428.X.

[20] Ellis RP, Layton TJ. Risk Selection and Risk Adjustment. Encycl Heal Econ 2014:289–97. https://doi.org/10.1016/B978-0-12-375678-7.00918-4.

[21] Centers for Medicare & Medicaid Services (CMS) H. Medicare Program; Medicare Shared Savings Program: Accountable Care Organizations. Fed Regist 2011.

[22] Sugiyama M. Introduction to Statistical Machine Learning. Introd to Stat Mach Learn 2015:1–498. https://doi.org/10.1016/C2014-0-01992-2.

[23] Marill KA. Advanced Statistics: Linear Regression, Part I: Simple Linear Regression. Acad Emerg Med 2004;11:87–93. https://doi.org/10.1197/J.AEM.2003.09.005.

[24] Marill KA. Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression. Acad Emerg Med 2004;11:94–102. https://doi.org/10.1197/J.AEM.2003.09.006.

[25] Tripepi G, Jager KJ, Dekker FW, Zoccali C. Linear and logistic regression analysis. Kidney Int 2008;73:806–10. https://doi.org/10.1038/SJ.KI.5002787.

[26] Tripepi G, Jager KJ, Dekker FW, Wanner C, Zoccali C. Measures of effect: Relative risks, odds ratios, risk difference, and "number needed to treat." Kidney Int 2007;72:789–91. https://doi.org/10.1038/SJ.KI.5002432.

[27] De'ath G, Fabricius KE. CLASSIFICATION AND REGRESSION TREES: A POWERFUL YET SIMPLE TECHNIQUE FOR ECOLOGICAL DATA ANALYSIS. Ecology 2000;81:3178–92. https://doi.org/10.1890/0012-9658.

[28] Loh WY. Classification and regression trees. Wiley Interdiscip Rev Data Min Knowl Discov 2011;1:14–23. https://doi.org/10.1002/WIDM.8.

[29] Breiman L. Random forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/A:1010933404324/METRICS.

[30] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55. https://doi.org/10.1093/BIOMET/70.1.41.

[31] Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Https://Doi-OrgEzproxy2UtwenteNl/101080/002731712011568786 2011;46:399–424. https://doi.org/10.1080/00273171.2011.568786.

[32] Lee J, Little TD. A practical guide to propensity score analysis for applied clinical research. Behav Res Ther 2017;98:76–90. https://doi.org/10.1016/J.BRAT.2017.01.005.

[33] Caliendo M, Kopeinig S. SOME PRACTICAL GUIDANCE FOR THE IMPLEMENTATION OF PROPENSITY SCORE MATCHING. J Econ Surv 2008;22:31–72. https://doi.org/10.1111/J.1467-6419.2007.00527.X.

[34] Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 2009;28:3083. https://doi.org/10.1002/SIM.3697.

[35] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33:1–22. https://doi.org/10.18637/JSS.V033.I01.

[36] Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 1970;12:55. https://doi.org/10.2307/1267351.

[37] Tibshirani R. Regression Shrinkage and Selection via The Lasso: A Retrospective. J R Stat Soc Ser B Stat Methodol 2011;73:273–82. https://doi.org/10.1111/J.1467-9868.2011.00771.X.

[38] Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net 2003.

[39] Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?-Arguments against avoiding RMSE in the literature. Geosci Model Dev 2014;7:1247–50. https://doi.org/10.5194/gmd-7-1247-2014.

[40] Van Kleef RC, Van Vliet RCJA, Van Rooijen EM. Diagnoses-based cost groups in the Dutch risk-equalization model: The effects of including outpatient diagnoses. Health Policy (New York) 2014;115:52–9. https://doi.org/10.1016/J.HEALTHPOL.2013.07.005.

[41] Lamers LM, Van Vliet RCJA. The Pharmacy-based Cost Group model: validating and adjusting the classification of medications for chronic conditions to the Dutch situation. Health Policy (New York) 2004;68:113–21. https://doi.org/10.1016/J.HEALTHPOL.2003.09.001.

[42] UC Business Analytics. Regression Trees · UC Business Analytics R Programming Guide n.d. https://uc-r.github.io/regression_trees (accessed June 17, 2023).

[43] Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci 2021;7:1–24. https://doi.org/10.7717/PEERJ-CS.623/SUPP-1.

[44] Venkatesh V, Bala H. Technology Acceptance Model 3 and a Research Agenda on Interventions Subject Areas: Design Characteristics, Interventions. Decis Sci 2008;39:273–315.

[45] Chaurasia A, Harel O. Using AIC in Multiple Linear Regression framework with Multiply Imputed Data. Health Serv Outcomes Res Methodol 2012;12:219. https://doi.org/10.1007/S10742-012-0088-8.

[46] Burnham K, Anderson D. Multimodel inference: Understanding AIC and BIC in model selection. Sociol Methods Res 2004:261–304. https://doi.org/10.1177/0049124104268644.

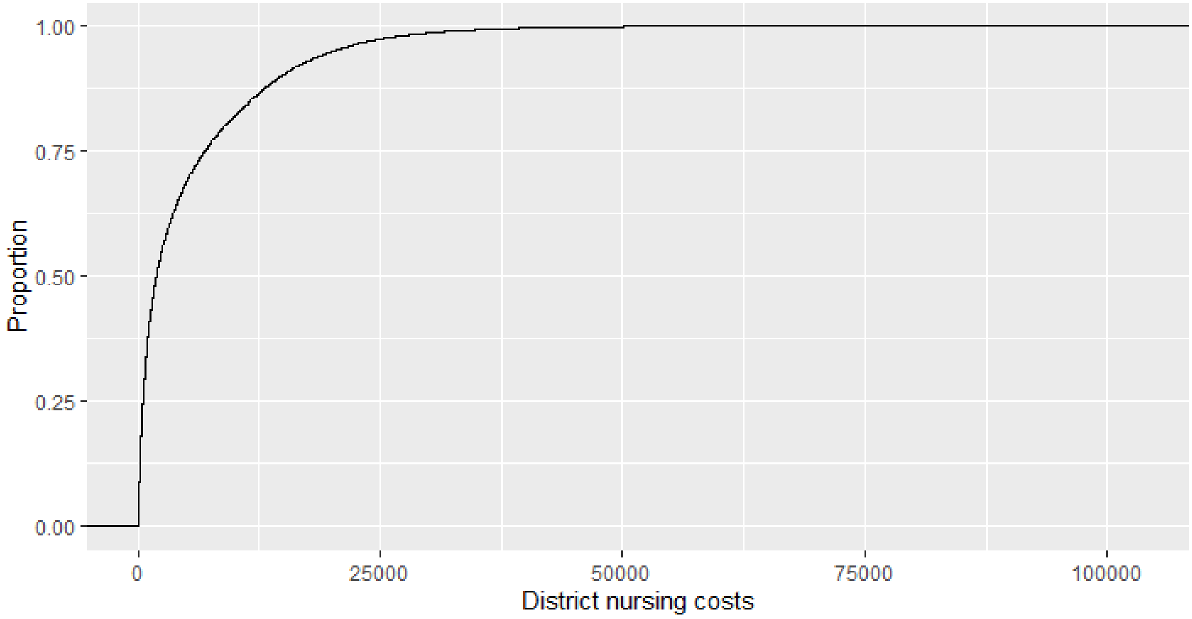
[47] Lee JD, Sun DL, Sun Y, Taylor JE. Exact post-selection inference, with application to the lasso. Https://DoiOrg/101214/15-AOS1371 2016;44:907–27. https://doi.org/10.1214/15-AOS1371.

[48] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc B 2006;68:49–67.

[49] Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection 1995.

Appendix

Appendix 1: the distribution of district nursing costs across the sample



Appendix 2: coefficients and p values of linear regression model applied to entire dataset

Variable	Coefficient	Std. Error	T value	P value	Significance
Intercept	2.315.170	1.281.227	1.807	0.070773	.
AANTAL_PERSONEN_ONDER_18_HH	129.755	145.331	0.893	0.371959	
AANTAL_PERSONEN_BOVEN_18_HH	-269.405	41.206	-6.538	6.33e-11	***
GESLACHTV	522.826	80.943	6.459	1.07e-10	***
BN_INKOMEN_OMS18.000-26.000euro	676.576	235.430	2.874	0.004059	**
BN_INKOMEN_OMS26.000-35.000euro	507.746	233.866	2.171	0.029931	*
BN_INKOMEN_OMS35.000-50.000euro	322.528	223.198	1.445	0.148459	
BN_INKOMEN_OMS50.000-75.000euro	241.585	210.999	1.145	0.252235	
BN_INKOMEN_OMS75.000-100.000euro	-111.557	215.204	-0.518	0.604199	
BN_INKOMEN_OMSMinderdan18.000euro	677.946	252.801	2.682	0.007328	**
BN_TWEEVERDIENERS_JN_OMSJa	251.223	173.541	1.448	0.147730	
BN_TWEEVERDIENERS_JN_OMSNee	194.461	152.229	1.277	0.201462	
BN_OPLEIDING_OMSHAVO/VWO/HBS	-99.282	268.429	-0.370	0.711485	
BN_OPLEIDING_OMSHBO-ofWO-bachelor	-143.523	231.726	-0.619	0.535681	
BN_OPLEIDING_OMSHBO-ofWO-master/MBA/Postdoctoraal	-172.137	355.245	-0.485	0.627993	
BN_OPLEIDING_OMSLBO/VMBO(kaderofberoep)/MBO1	-38.542	108.955	-0.354	0.723537	
BN_OPLEIDING_OMSMAVO/MULO/VMBO(theoretischofgemengd)	6.720	156.769	0.043	0.965807	
BN_OPLEIDING_OMSMBO(2,3of4)	62.208	148.539	0.419	0.675365	
BN_ZORGCLIENT_TYPE_OMSEigenzinnigezorgclient	-103.195	245.385	-0.421	0.674092	
BN_ZORGCLIENT_TYPE_OMSGemaksgerichtezorgclient	337.319	141.348	2.386	0.017019	*

BN_ZORGCLIENT_TYPE_OMSKwaliteitsgerichtezorgclient	8.401	212.319	0.040	0.968439	
BN_ZORGCLIENT_TYPE_OMSLuxegerichtezorgclient	-241.603	182.602	-1.323	0.185807	
BN_ZORGCLIENT_TYPE_OMSMaatschappijkritischezorgclient	117.012	194.507	0.602	0.547457	
BN_ZORGCLIENT_TYPE_OMSResultaatgerichtezorgclient	280.191	169.347	1.655	0.098030	.
BN_ZORGCLIENT_TYPE_OMSVolgzamezorgclient	158.849	108.148	1.469	0.141896	
Overleden_indicator_halfjaar	1679.263	173.396	9.685	<2E-16	***
MSZ_2021	-408.643	163.755	-2.495	0.012585	*
DKG_01	213.447	139.406	1.531	0.125750	
DKG_11	148.393	142.878	1.039	0.299000	
DKG_21	-324.090	147.807	-2.193	0.028340	*
DKG_31	103.797	141.479	0.734	0.463163	
DKG_41	444.658	169.185	2.628	0.008587	**
DKG_51	438.355	185.255	2.366	0.017976	*
DKG_61	-77.755	157.567	-0.493	0.621682	
DKG_71	-129.327	393.063	-0.329	0.742140	
DKG_81	2426.920	401.896	6.039	1.57e-09	***
DKG_91	2098.884	632.750	3.317	0.000911	***
DKG_101	-511.476	328.790	-1.556	0.119806	
DKG_111	415.046	359.153	1.156	0.247844	
DKG_121	6590.250	696.555	9.461	<2,00E-16	***
DKG_131	3114.562	1.987.833	1.567	0.117169	
DKG_141	837.907	523.746	1.600	0.109645	
DKG_151	-950.090	687.565	-1.382	0.167037	
FKG_01	17.531	125.536	0.140	0.888939	
FKG_11	-2.994	162.887	-0.018	0.985333	
FKG_21	698.334	174.799	3.995	6.48e-05	***
FKG_31	1024.848	158.535	6.464	1.03e-10	***
FKG_41	1266.622	268.460	4.718	2.39e-06	***
FKG_51	1506.594	275.592	5.467	4.62e-08	***
FKG_61	647.580	147.622	4.387	1.15e-05	***
FKG_71	117.684	478.239	0.246	0.805623	
FKG_81	1768.096	293.400	6.026	1.70e-09	***
FKG_91	1253.025	121.668	10.299	<2.00E-16	***
FKG_101	1149.887	198.053	5.806	6.46e-09	***
FKG_111	1513.625	266.851	5.672	1.42e-08	***
FKG_121	727.621	292.780	2.485	0.012953	*
FKG_131	714.761	150.338	4.754	2.00e-06	***

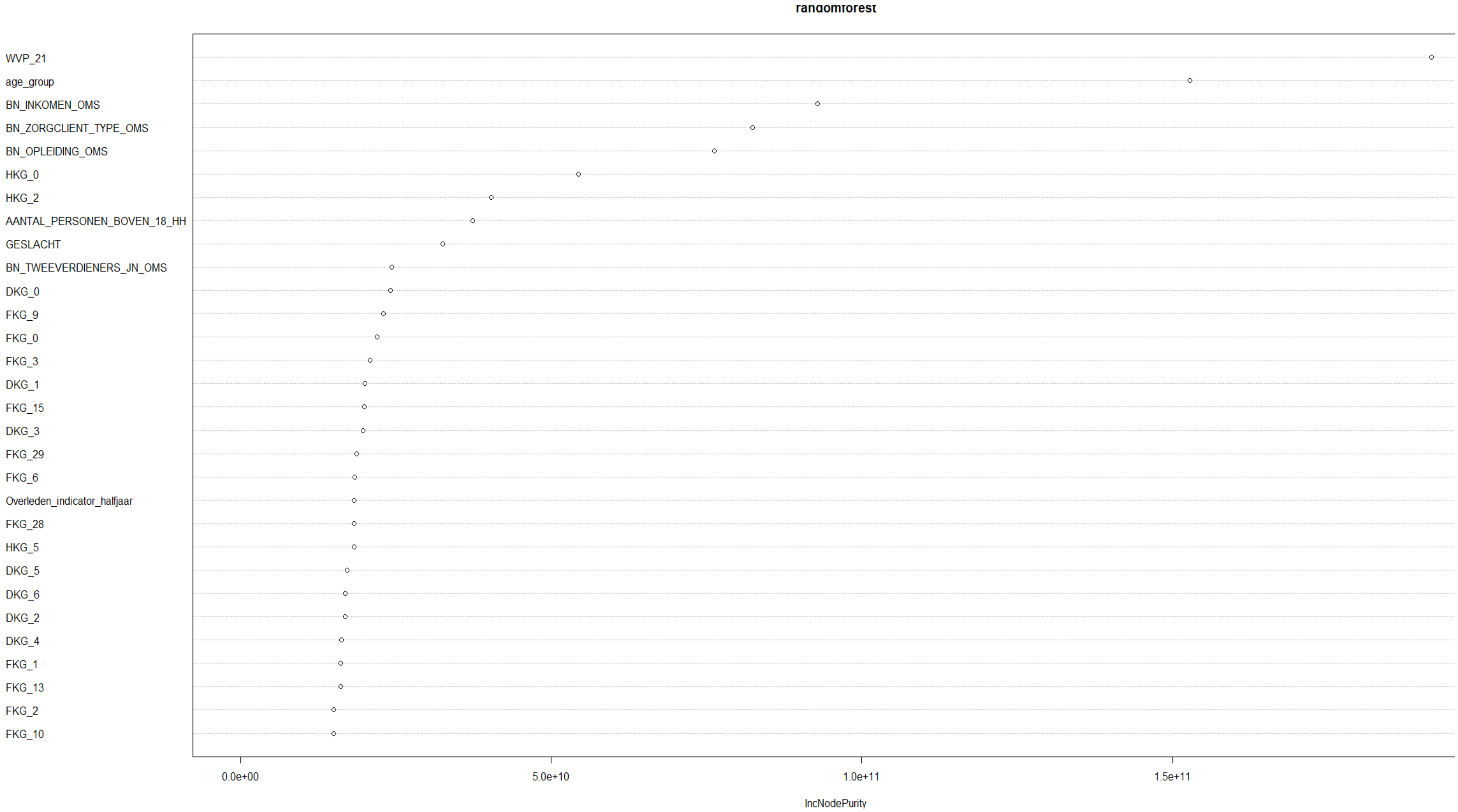
FKG_141	1670.580	324.467	5.149	2.64e-07	***
FKG_151	2335.571	153.800	15.186	<2.00E-16	***
FKG_161	-930.008	658.550	-1.412	0.157900	
FKG_171	2613.135	2483.279	1.052	0.292674	
FKG_181	3468.303	640.457	5.415	6.16e-08	***
FKG_191	1600.704	1211.271	1.322	0.186342	
FKG_201	-2942.256	1150.551	-2.557	0.010555	*
FKG_211	1143.619	698.282	1.638	0.101482	
FKG_221	-647.302	491.700	-1.316	0.188031	
FKG_231	398.469	318.266	1.252	0.210579	
FKG_241	29.242	451.473	0.065	0.948357	
FKG_251	535.652	553.237	0.968	0.332945	
FKG_261	764.525	1269.763	0.602	0.547111	
FKG_271	1191.735	1352.885	0.881	0.378387	
FKG_281	417.371	160.048	2.608	0.009117	**
FKG_291	511.071	156.237	3.271	0.001072	**
FKG_301	-1346.379	1442.044	-0.934	0.350487	
FKG_311	-1277.195	304.754	-4.191	2.79e-05	***
FKG_321	191.553	1292.772	0.148	0.882208	
FKG_331	-812.715	340.214	-2.389	0.016908	*
FKG_341	-290.828	1708.678	-0.170	0.864849	
FKG_351	-5814.540	6599.550	-0.881	0.378297	
HKG_01	-568.945	157.518	-3.612	0.000304	***
HKG_11	-332.935	226.083	-1.473	0.140862	
HKG_21	2676.705	164.887	16.234	<2.00E-16	***
HKG_31	1367.671	269.634	5.072	3.95e-07	***
HKG_41	219.503	374.774	0.586	0.558085	
HKG_51	2431.729	218.009	11.154	<2.00E-16	***
HKG_61	636.548	266.543	2.388	0.016939	*
HKG_71	1522.293	343.817	4.428	9.56e-06	***
HKG_81	998.776	546.687	1.827	0.067715	.
HKG_91	-1347.692	2509.642	-0.537	0.591268	
HKG_101	-533.456	607.582	-0.878	0.379951	
WVP_21	3990.411	81.846	48.755	<2.00E-16	***
age_group5	-1646.455	1416.606	-1.162	0.245142	
age_group6	-2320.050	1388.418	-1.671	0.094731	.
age_group7	-2263.375	1330.672	-1.701	0.088967	.
age_group8	-1373.281	1302.744	-1.054	0.291825	
age_group9	-1821.336	1278.490	-1.425	0.154283	
age_group10	-2114.259	1253.294	-1.687	0.091621	.
age_group11	-2004.398	1238.346	-1.619	0.105542	
age_group12	-1562.200	1233.537	-1.266	0.205365	
age_group13	-1697.675	1229.365	-1.381	0.167309	
age_group14	-1699.545	1226.817	-1.385	0.165962	
age_group15	-1456.758	1224.813	-1.189	0.234303	
age_group16	-1133.847	1224.577	-0.926	0.354501	

age_group17	-778.824	1224.295	-0.636	0.524689	
age_group18	109.771	1225.918	0.090	0.928652	
age_group19	1302.635	1232.627	1.057	0.290613	
age_group20	3913.442	1290.278	3.033	0.002423	**
age_group21	13166.701	1937.960	6.794	1.11e-11	***

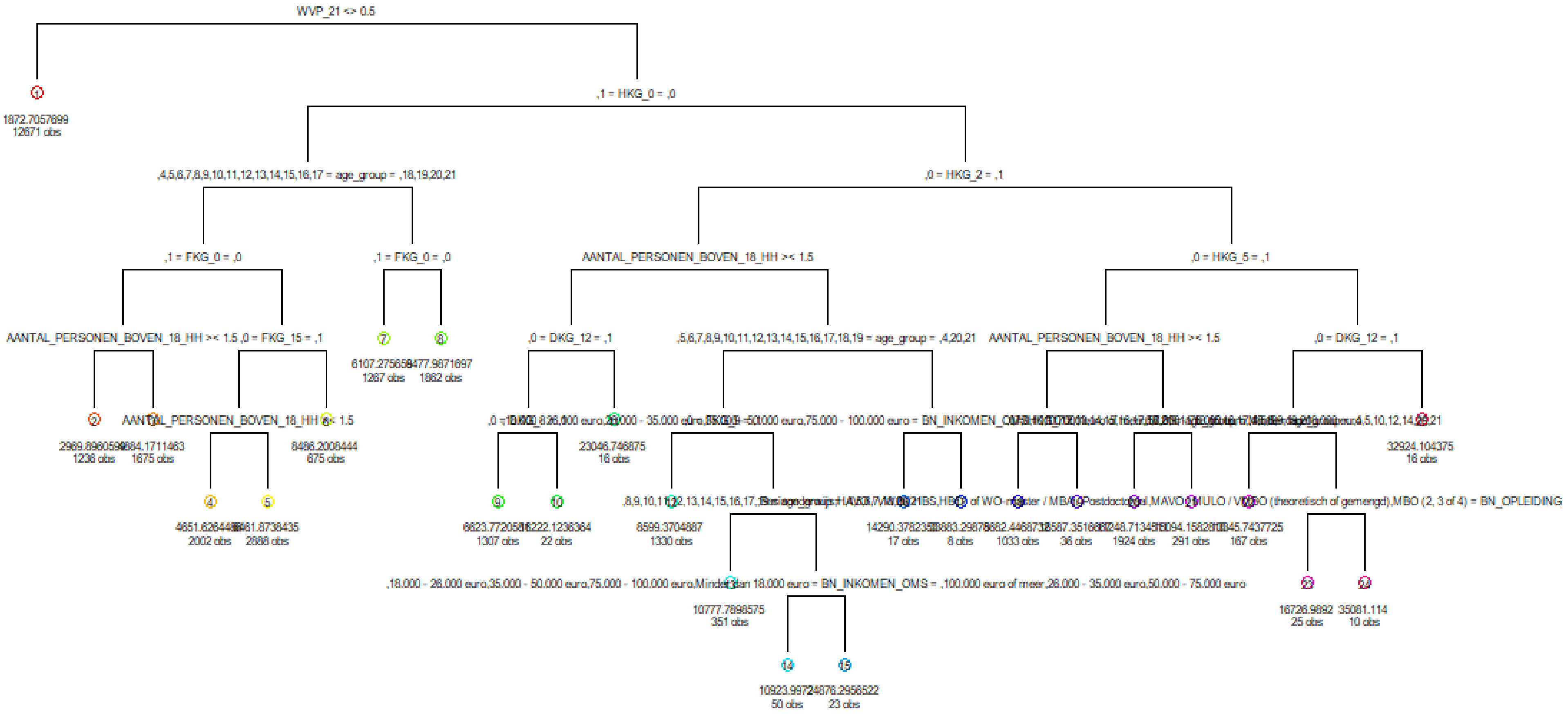
Appendix 3: coefficients of elastic net regression model applied to entire dataset

Variable	**Coefficient**
Intercept	2925
Age_group 21	6163
WVP_21	4106
DKG_12	4060
Age_group 20	2842
HKG_2	2097
age_group 19	1532
HKG_5	1529
FKG_15	1275
Overleden_indicator_halfjaar	845
FKG_9	762
Age_group 18	690
FKG_18	587
FKG_11	359
GESLACHTV	234
DKG_8	225
BN_INKOMEN_OMS 18.000-26.000 euro	65
HGK_3	38
FKG_33	-43
BN_ZORGCLIENT_TYPE_OMS Luxegerichte zorgclient	-70
BN_INKOMEN_OMS 75.000 – 100.000 euro	-98
AANTAL_PERSONEN_BOVEN_18_HH	-207
FKG_0	-599
HKG_0	-969

Appendix 4: variable importance plot of random forest applied to entire dataset



Appendix 5: regression tree applied to the whole dataset



Appendix 6: PowerPoint presentation, as presented to health care purchasers at Menzis

Presentatie
zorginkoop.pptx