Exploring the Predictive Power of Interaction Terms in a Sophisticated Risk Equalization Model using Regression Trees

Published in:
Health Economics

Publication status and date:
Published: 01/01/2018

DOI (link to publisher):
10.1002/hec.3523

Document Version
Publisher's PDF, also known as Version of record

Document License/Available under:
Article 25fa Dutch Copyright Act

Citation for the published version (APA):
Veen, SHCM., van Kleef, R., van Vliet, R., & van de Ven, W. (2018). Exploring the Predictive Power of Interaction Terms in a Sophisticated Risk Equalization Model using Regression Trees. *Health Economics*, *27*, e1-e12. https://doi.org/10.1002/hec.3523

Received: 17 December 2014 | Revised: 29 November 2016 | Accepted: 5 April 2017
DOI: 10.1002/hec.3523

RESEARCH ARTICLE

WILEY

Exploring the predictive power of interaction terms in a sophisticated risk equalization model using regression trees

S. H. C. M. van Veen | R. C. van Kleef | W. P. M. M. van de Ven | R. C. J. A. van Vliet

Institute of Health Policy and Management, Erasmus University Rotterdam, Rotterdam, the Netherlands

Correspondence
S.H.C.M. (Suzanne) van Veen, Institute of Health Policy and Management, Erasmus University Rotterdam, DR Rotterdam, the Netherlands.
Email: vanveen@bmg.eur.nl

JEL Classification: C38; C49; I13; I18

Abstract

This study explores the predictive power of interaction terms between the risk adjusters in the Dutch risk equalization (RE) model of 2014. Due to the sophistication of this RE-model and the complexity of the associations in the dataset (N = ~16.7 million), there are theoretically more than a million interaction terms. We used regression tree modelling, which has been applied rarely within the field of RE, to identify interaction terms that statistically significantly explain variation in observed expenses that is not already explained by the risk adjusters in this RE-model. The interaction terms identified were used as additional risk adjusters in the RE-model. We found evidence that interaction terms can improve the prediction of expenses overall and for specific groups in the population. However, the prediction of expenses for some other selective groups may deteriorate. Thus, interactions can reduce financial incentives for risk selection for some groups but may increase them for others. Furthermore, because regression trees are not robust, additional criteria are needed to decide which interaction terms should be used in practice. These criteria could be the right incentive structure for risk selection and efficiency or the opinion of medical experts.

KEYWORDS

interaction terms, regression trees, risk equalization

1 | INTRODUCTION

1.1 | Background

Risk equalization (RE) models are widely used—Belgium, Germany, Israel, the Netherlands, Switzerland, and the U.S. health care exchanges—for calculating risk-adjusted payments to health insurers in order to compensate them for predictable differences in individuals' health care expenses (Ash, Porell, Gruenberg, Sawitz, & Beiser, 1989; Kautter, Pope, & Keenan, 2014; van de Ven, van de Beck, van de Voorde, Wasem, & Zmora, 2007). In the presence of premium regulation, as is the case in all of the aforementioned countries, the goal of RE is to mitigate financial incentives for risk selection and thereby achieve a level playing field for health insurers.

During the past three decades, many studies have investigated the predictive power of new types of risk adjusters in RE-models, for example, demographic risk adjusters, morbidity-based risk adjusters relying on pharmaceutical or diagnostic information, or cost-based risk adjusters (e.g., Adams, Bronstein, & Raskind-Hood, 2002; Fishman et al., 2003; Hughes et al., 2004; Kronick, Gilmer, Dreyfus, & Lee, 2000; Lamers & van Vliet, 2003; Pope, Ellis, Ash, Liu, et al., 2000). Relatively little systematic attention, however, has been paid to the predictive power of interaction terms—hereafter "interactions"—between the risk classes of the risk adjusters in the RE-model. The motive for using interactions is that the risk classes in the RE-model may be heterogeneous with respect to expected health care expenses. To a certain extent, this heterogeneity may be explained by interactions between the risk classes. Some studies have demonstrated that interactions can improve models' predictive performance and thus can mitigate financial incentives for risk selection (Buchner, Wasem, & Schillo, 2015; Pope, Ellis, Ash, Ayanian, et al., 2000; Pope et al., 2004; Robinson, 2008; Zhao et al., 2001, 2005).

Interactions have been applied moderately in existing RE-models. Some RE-models, such as those used in Belgium, Germany, the Netherlands, and Switzerland, include some first-order interactions, for example, age interacted with gender. However, to our knowledge, none of the existing RE-models use higher order interactions or interactions among morbidity-based risk adjusters, while these may be especially useful to predict expenses for selective groups adequately, such as individuals with multiple morbidities (Pope et al., 2004). Several studies have shown that even morbidity-based RE-models do not adequately predict expenses for specific patient groups (e.g., Behrend et al., 2007; Payne, Cebul, Singer, Krishnaswamy, & Gharrity, 2000; van Kleef, van Vliet, & van de Ven, 2014). In the presence of premium regulation, underprediction and overprediction of expenses for such selective groups provide insurers with financial incentives for risk selection, which is a potential threat to solidarity, efficiency, and quality of care (Newhouse, 1996; van de Ven & Ellis, 2000).

1.2 | Study objective and contribution

This study explores the predictive power of interactions in the Dutch RE-model of 2014. This RE-model includes four morbidity-based risk adjusters, namely, a risk adjuster for prior use of specific prescription drugs in terms of pharmaceutical cost groups (PCGs), prior hospitalization in terms of diagnostic cost groups (DCGs), prior use of certain durable medical equipment (DME-groups), and multiple-year high costs over the three preceding years (MHC-groups). Due to the sophistication of this morbidity-based RE-model and the complexity of the associations in datasets with millions of observations and many relevant risk factors, which are common in the field of RE, there are theoretically more than a million interactions. Probably not all of these interactions are relevant from a statistical point of view. We use regression trees to automatically identify relevant interactions, preventing exhaustive hand searches. With hand searches, it is practically impossible to find all interactions that are statistically relevant, some of which may be unexpected because of complex, unknown relationships. Regression trees in this study identify interactions that statistically significantly explain variation in observed expenses that is not already explained by the risk adjusters in the Dutch RE-model of 2014. The interactions identified are then used as additional risk adjusters. By comparing the predictive performance of the extended RE-models with the predictive performance of the Dutch RE-model, we conclude to what extent interactions improve model's predictive performance; we did not aim to conclude which interactions should be used.

Although regression trees have been used extensively in various scientific fields, they have been applied rarely within the context of RE. To our knowledge, Robinson (2008) was the first who used regression trees to predict individual health care expenses. For the purpose of our study, we do not use regression trees to predict expenses per se, but we used them to identify possible interactions explaining the residuals of the RE-model under study. In a second step, the interactions identified are used as additional risk adjusters in the RE-model, given the ordinary least squares (OLS) model as used in practice. Buchner et al. (2015) were the first to apply such a "two-step approach." Though our study objective is similar to theirs, we extend on their work by the following three methodological improvements. First, in developing the regression trees, we use a more efficient definition of the target variable (i.e., the dependent variable in the regression tree; see Section 2). Second, instead of estimating a single tree, several trees are estimated to test the robustness with respect to the identification of interactions. Third, owing to the availability of external information, that is, information that is not explicitly used in developing the RE-model, we were able to assess models' performance for selective groups of individuals in order to measure financial incentives for risk selection with and without the interactions included in the RE-model.[1]

Extending RE-models with interactions may improve predictive performance and thereby mitigate financial incentives for risk selection. Interactions can be especially useful in modelling highly skewed expenses, because such types of expenses require accounting for nonlinearities in the data. Expenses of several types of services, such as hospitalization or long-term care, typically have skewed distributions. To predict these cost types adequately, interactions may be relevant.

Section 2 describes the data and methodology. Results are presented in Section 3. Section 4 concludes and discusses the results.

[1]Using evaluation-groups that are similar to those explicitly included in the RE-model would overestimate model's performance (van Veen et al., 2015).

2 | DATA AND METHODOLOGY

2.1 | Administrative data and health survey data

Dutch administrative data of 2011 was used to develop the regression trees and estimate the RE-models. Information on total health care expenses, age, gender, source of income, socioeconomic status, region, PCGs, DCGs, DME-groups, and MHC-groups were available for each individual (N = ~16.7 million). Total expenses included all costs covered by the Dutch basic benefit package, except mental health care services.[2,3]

A Dutch health survey, "Gecon," was used to assess models' performance at selective groups. This survey is conducted each year on a representative sample of the Dutch population by "Statistics Netherlands."[4] The survey is targeted at households. Individuals in mental health care institutions and nursing homes are excluded. The survey results from 2010 were merged at the individual level with the administrative dataset using an anonymous identification variable (N = 16,141).[5] Information on self-reported health status and health care utilization were used to define the evaluation groups.

2.2 | Regression trees

2.2.1 | How do regression trees work?

Regression trees, developed by Breiman and colleagues in 1984, are nonparametric techniques and belong to the family of "classification and regression trees" (CART). Nonparametric techniques in contrast to parametric techniques, such as regression modeling, do not require prior knowledge about the functional relationship between the variables in the model (e.g., linear, logistic, or logarithmic). These techniques are data-driven, implying that any functional relationship can be estimated that fits the data best. Such techniques can lead to unexpected relationships that were first unknown, providing new valuable insights, for example, specific combinations of patient characteristics that lead to large residuals under the current RE-model. Within the family of CART, regression trees are used when the target variable is a continuous variable and classification procedures are used when the target variable is a categorical variable. Both techniques partition the data into smaller and smaller pieces, representing subgroups with similar characteristics. They do not use a prespecified functional relationship, but they search for complex associations based on some user-defined stopping rules for splitting the data. Below, we will briefly explain in more detail how regression trees work. For a thorough discussion of the technical details of CART see Breiman, Friedman, Stone, & Olshen, 1984; Hastie, Tibshirani, & Friedman, 2009; and Strobl, Malley, & Tutz, 2009.

Developing a regression tree starts with "growing" the total tree, which basically means that the data are recursively partitioned into subgroups (Berk, 2006; Hastie et al., 2009; Sarma, 2007). As a first step, the tree automatically searches through all key variables (i.e., independent variables) one by one and chooses the best split. The best split is the split that results into two or more subgroups that are as homogeneous as possible with respect to the target variable (within-group variation) and that are maximally differentiated from the other subgroups in terms of the target variable (between-group variation). Whether the data at each step is split in two subgroups, that is, a binary tree, or in more subgroups, that is, a multiway tree, depends on a user-defined model parameter. At all next steps, the tree searches the best split, given ancestor split(s). This process continues until prespecified stopping rules are met, such as the statistical significance of the F statistic[6] or a minimum number of observations per subgroup. In Breimans' terminology, the hierarchy of groups is called a *tree*, the intermediate subgroups are called *nodes*, and the final subgroups are called *leafs*. The leafs are mutually exclusive and define the interactions.

When growing the total tree, overfitting may occur, meaning that the leafs describe noise rather than the underlying relationship in the data (Berk, 2006, Hastie et al., 2009). To prevent overfitting, the trees can be pruned. "Pruning" is a process that

[2]The basic benefit package included expenses related to hospital care, primary care, paramedical care, pharmaceuticals, durable medical equipment, medical transport, dental care, obstetrical care, maternity care, and mental health care services. Expenses for mental health care services have been excluded, because a separate RE-model with specific mental health care risk adjusters is applied for these expenses.

[3]Total expenses were annualized and weighted by the fraction of the year the individual was enrolled, for example, an individual who was enrolled for 6 months and had €500 expenses was given a weight of 0.5 and €1,000 annual expenses.

[4]"Statistics Netherlands" is an autonomous Dutch agency that collects data and publishes statistics to be used by policymakers and researchers.

[5]The survey results from 2010 (and not 2011) were used, because we aimed to investigate the extent to which the RE-models adjust for predictable differences in individuals' health care expenses, using information known prior to the estimation year (2011).

[6]The F statistic is determined by the average differences in residual expenses across groups and the size of the groups. Consequently, small groups with high residual expenses may be overlooked. In the context of RE, these small groups can be of particular interest. It is beyond the scope of this study to develop a tree where the splits are solely based on average differences in residual expenses. For further research, it would be interesting to compare the results of such a tree with our results.

sequentially removes leafs from the bottom of the total tree and selects the subtree with the highest accuracy (Sarma, 2007), meaning the one with the minimum weighted average of the average squared errors of all leafs of the subtree (Sarma, 2007).

2.2.2 | Data preparations

The total administrative dataset was split in three samples: *sample 1* for growing the trees, *sample 2* for pruning the trees and estimating the coefficients of the RE-models, and *sample 3* for predicting expenses by the RE-models and assessing models' performance. To prevent overfitting of the trees, we used one sample for growing the trees and another sample for pruning the trees. The leafs of the pruned trees were used for defining the interactions. The third sample was needed to prevent overestimation of the predictive performance of the extended RE-models.

To assign individuals to one of the samples, all respondents to the survey were first assigned to sample 3 in order to fully exploit this dataset for model evaluation. After this, all remaining individuals were randomly assigned to one of the samples in such a way that sample 1 contained ~50% of the total observations and sample 2 and sample 3 both contained ~25% of the total observations. Consistent with other studies, for example, Robinson, 2008; Hastie et al., 2009, more data were reserved for growing the tree than for pruning or evaluation, because this generally results in more stable estimates of the tree (Sarma, 2007). We used this split sampling approach to follow the standard procedure for addressing overfitting, even though overfitting turns out to be an unimportant problem in our study due to the large sample size (Ellis & Mookim, 2009, Hastie et al., 2009).[7]

In the total administrative dataset, mean observed expenses were €1,785 and mean residual expenses were €0[8] (Table 1). Average age was 40 years, 49.3% of the individuals were male, 17.3% were classified into a PCG, with 3.5% having more than one PCG, 8.7% were classified into a DCG, 5.8% were classified into an MHC-group, and 0.8% were classified into a DME-group. Combining these risk adjusters, 22% were classified into a PCG, DCG, MHC-group, and/or DME-group. As shown in Table 1, our split sampling procedure did not yield bias in the representativeness of the administrative samples.

To develop the trees, all individuals who were not enrolled the full year were excluded: ~3% in sample 1 and sample 2. The reason for this exclusion was that SAS© Enterprise Miner 12.1 did not offer a satisfactory method for incorporating weights in developing the trees. With our approach, deceased persons and most of the newborns, who generally have above-average expenses, were excluded from sample 1 and sample 2 that were used for developing the trees. Consequently, some interesting interactions for these groups may be overlooked. However, for the estimation and evaluation of the RE-models, total sample 2 and total sample 3 were used. Individuals who were enrolled for a part of the year were classified to one of the interaction-groups according to their risk characteristics.

After exclusion of those who were not enrolled, the full year samples 1 and 2 contained ~8,1 million and ~4,1 million individuals, respectively. Descriptive statistics on costs and prevalence of the risk adjusters were quite similar to those for the total administrative dataset after exclusion of the partially enrolled.[9] Average total costs were €1,691 and €1,692 for samples 1 and 2, respectively. Average age was 40 years, 49.3% of the individuals were male, 17.3% were classified into a PCG, with 3.4% having more than one PCG, 8.6% were classified into a DCG, 5.7% were classified into an MHC-group, and 0.8% were classified into a DME-group. Combining these risk adjusters, 22% were classified into a PCG, DCG, MHC-group, and/or DME-group.

2.2.3 | Regression tree model specifications

The target variable in the trees was residual expenses[10,11] defined as observed expenses minus predicted expenses by the Dutch RE-model of 2014.[12] With our definition of the target variable, the tree finds interactions that explain variation in observed

[7]To test whether overfitting occurs, we compared the predictive power of the leafs of a total tree to the leafs of a pruned tree, ceteris paribus. The motive for this test is that removed leafs may describe patterns in the sample that is used for growing the tree well but may not do so in another sample. The outcomes of the test show that the removed leafs do not have high predictive power: The R^2 and the CPM of the RE-model with the interactions of a pruned tree versus those of a total tree dropped 0.01 percentage point and the MAPE remained the same. This finding shows that overfitting is not a serious problem in our study for the purpose of exploring to what extent model's performance can be improved. A reason for this is the usage of large datasets with millions of observations.

[8]This is the result of estimating the RE-model with OLS-methods on the total administrative dataset.

[9]The descriptive statistics can be provided upon request from the first author.

[10]Residual health care expenses have typically long tails. Although transformation can increase model fit, we did not do this, because we endeavored identifying interactions that fit well residual expenses expressed in Euros, and not in another unit of measurement, for example, log-Euros.

[11]Residual expenses as a dependent variable is only used to identify interactions. To estimate the coefficients of the risk adjusters in the RE-models with and without the interactions, observed expenses is the dependent variable.

[12]This model was estimated on the total administrative dataset. The dependent variable was observed expenses and the independent variables were the eight risk adjusters (Table 3). The model used a constant and a weight for the enrollment period.

TABLE 1 Descriptive statistics of the administrative dataset from the Dutch population of insured in 2011 (N = ~16.7 million), the three samples of this administrative dataset, and the health survey sample

	Total administrative dataset	Samples of the administrative dataset: Sample 1	Sample 2	Sample 3	Health survey sample
N (individuals)	16,688,961	8,327,580	4,247,646	4,113,735	16,141
N (insured years)	16,438,958	8,201,696	4,184,047	4,053,215	16,067
Mean total observed health care expenses in Euros (std.)[a, b]	1,785 (5,978)	1,783 (5,944)	1,785 (6,131)	1,787 (5,885)	1,766 (5,364)
Median total observed health care expenses in Euros	445	445	445	446	444
Mean age in years (std.)[b]	40.0 (22.9)	40.1 (22.9)	40.1 (22.9)	40.1 (22.9)	39.7 (23.0)
Proportion male	0.493	0.493	0.493	0.493	0.486
Proportion classified in a PCG	0.173	0.173	0.173	0.173	0.172
Proportion classified in multiple PCGs	0.035	0.035	0.035	0.035	0.035
Proportion classified in a DCG[c]	0.087	0.086	0.086	0.087	0.087
Proportion classified in an MHC-group	0.058	0.058	0.058	0.058	0.059
Proportion classified in a DME-group	0.008	0.008	0.008	0.008	0.009
Proportion classified in a PCG, DCG, MHC-group, and/or DME-group	0.220	0.220	0.220	0.221	0.220

Note. All statistics are weighted for the enrolment period of individuals. DCG = diagnostic cost group; DME-group = durable medical equipment group; MHC-group = multiple-year high cost group; PCG = pharmaceutical cost group

[a]Observed expenses are annualized and weighted for the enrolment period in 2011. All expenses are rounded to the nearest Euro.

[b]To calculate the standard deviation, the sum of the weights minus one is used as the variance divisor.

[c]Individuals can be classified in only one DCG, the one with the highest follow-up costs.

expenses that is not already explained by the risk adjusters in this RE-model. This definition is more efficient than using observed expenses as the target variable, since then, the tree finds interactions that explain variation in observed expenses, but not all of them may have significant predictive power when the additive effects of the risk adjusters are incorporated.[13]

The key variables in the trees were the risk adjusters in the Dutch RE-model of 2014. Table 2 provides the definition of these risk adjusters. For the PCGs, binary variables were used instead of a categorical variable, because individuals can be classified into multiple PCGs. For each of the other risk adjusters, the risk classes are mutually exclusive. We did not use binary variables for all risk classes, because binary variables have less flexibility in defining splits than categorical variables; that is, splits by a binary variable can only be based on being or not being in a risk class, while a split by a categorical variable can be based on multiple risk classes. The risk adjusters are taken as given. We did not examine finer categories of them. Furthermore, the risk adjusters in the Dutch RE-model are chosen to achieve statistical as well as political and practical objectives, such as appropriateness of incentives for efficiency and availability of data. We purely focus on statistical criteria, given the risk adjusters as used in practice.

Five trees were developed to examine the extent to which the specification of the tree influences the identification of interactions. *Tree 1* was grown on sample 1 and pruned on sample 2. To indicate the influence of the type of sample used for growing and pruning the tree, without being influenced by the sample size (Gail, Krickeberg, Samet, Tsiatis, & Wong, 2009; Hastie et al., 2009; Last, Maimon, & Minkov, 2002), we grew *Tree 2* on sample 2 and pruned it on a random half of sample 1 and we grew *Tree 3* on the same random half of sample 1 and pruned it on sample 2. Comparing Tree 3 to Tree 1 indicates the influence of the sample size that was used for growing the tree (Oates & Jensen, 1997). To investigate the sensitivity of the tree to outliers in the target variable (Berk, 2006), *Tree 4* used residual expenses that were truncated at €−50,000 and € + 50,000. In sample 1 before applying truncation, 0.08% of the individuals had residual expenses above the cost caps, varying from €−93,253 to € + 2,059,572. *Tree 5* used a maximum depth of three levels, causing the tree to not have higher than third-order terms. All other trees have no restrictions regarding the depth of the tree. The reason to develop Tree 5 is that trees can easily become too complex when there are many variables and observations, resulting in higher order interactions that may be too complex

[13]An additional test, where we estimated a tree with observed expenses as the target variable, ceteris paribus, showed that more interactions were identified than a tree with residual expenses as the target variable: 327 versus 128 interactions. However, these 327 interactions have about a similar predictive power: The R^2 of an RE-model with these 327 interactions is 0.06 percentage point higher than an RE-model with 128 interactions, the CPM is 0.13 percentage point higher, and the MAPE is 2 Euros lower. Our approach of using residual expenses as the target variable is more effective than using observed expenses, because the additional predictive power per interaction term is higher than using the alternative approach.

TABLE 2 Definition of the risk adjusters in the Dutch RE-model of 2014

Risk adjusters and definitions
Age * gender (40 risk classes)
20 age classes for males and 20 age classes for females, with age in 5-year classes, starting from 0 year, 1–4 years, 5–9 years, 10–14 years, 15–17 years, 18–24 years up to an age of 90. Individuals older than 90 years are included in a separate risk class.
Source of income * Age (18 risk classes)
4 categories of source of income (disability benefits, social security benefits, self-employed, and others) are interacted with 4 classes of age (18–34 years, 35–44 years, 45–54 years, and 55–64 years). There is a separate class for individuals younger than 18 years or older than 64 years and a separate class for students with an age 18–34 years.
Region (10 risk classes)
10 risk classes, of which each class consists of a cluster of—not necessarily adjacent—zip codes areas. The clustering of the zip codes is based on several risk characteristics of the zip codes areas, such as the percentage of nonwestern immigrants, percentage of one-person households, distance to the general practitioner, distance to a hospital, and degree of urbanization.
Socioeconomic status * Age (12 risk classes)
4 socioeconomic classes: SES 0 is for individuals living in a home address with more than 15 persons (i.e., residents homes), SES 1 is for individuals in a household with an income in the lowest three deciles of the income distribution, SES 2 is for individuals in a household with an income in the following four deciles of the income distribution, and SES 3 is for individuals in a household with an income in the highest three deciles of the income distribution, interacted with 3 age classes of 0–17 years, 18–64 years, and individuals older than 64 years.
Pharmaceutical cost groups (PCGs, 24 risk classes)
Individuals are assigned to a PCG when they have used more than 180 daily dosages of a specific prescribed drug in the previous year. Individuals who did not use prescribed drugs in the previous year or individuals who did not use more than 180 daily dosages of that drug in the previous year were classified in PCG 0. Individuals can be classified into multiple PCGs, with some restrictions on combinations of PCGs, e.g., there are multiple PCGs for use of insulin, but individuals can be classified into only one of these PCGs.
Diagnostic cost groups (DCGs, 16 risk classes)
Individuals are assigned to a DCG when they have had a hospital admission in the previous year for specific medical diagnoses. Individuals with no hospital admission were classified in DCG 0. Individuals can be classified into only one DCG, the one with the highest follow-up costs.
Multiple-year high cost groups (MHC-groups, 7 risk classes)
Individuals are classified in a risk class when they belong three consecutive years to the top 15%, top 10%, top 7%, top 4%, or top 1.5% of the cost distribution in each year, or when they belong two consecutive years to the top 10% of the cost distribution in each year. Individuals who did not have high expenses in multiple years are classified in a separate risk class (MHC 0).
Durable medical equipment groups (DME-groups, 5 risk classes)
Individuals are classified to a risk class when they have used certain durable medical equipment: DME 1 is for those individuals who used insulin pumps, DME 2 is for those who used a catheter, DME 3 is for those who used a colostomy, and DME 4 is for those who used a trachea colostomy. Individuals who did not have used durable medical equipment are classified in a separate risk class (DME 0).

to be used in practice. We choose to restrict the tree to a maximum depth of 3 because a depth of 2 may lead to only a few interactions and a higher depth may lead to many complex interactions. Note that there is a wide gap between a tree restricted to interactions with a maximum depth of 3 and an unrestricted tree. However, comparing the other trees to tree 5 indicates the degree of what is possible between using a relatively simple tree and a complex tree with no restrictions on depth.

Next to the aforementioned model specifications, it was required to specify some additional parameters.[14] For ease of comparability, these additional specifications were equivalent across the five aforementioned trees. First, a level of 0.05 was used to test the statistical significance of splits. Second, the p values that were used for testing the statistical significance of splits were adjusted using the Bonferroni- and depth-correction (Sarma, 2007). These corrections make the statistical tests more stringent. Third, for reasons of stability, a minimum leaf size was specified. To set this rule, the number of individuals in the smallest risk class in the Dutch RE-model of 2014 was used, because this appears to be acceptable. The smallest risk class in the total administrative dataset was a DCG for hemophilia, leading to a minimum leaf size of 862 individuals for Trees 1, 4, and 5 (sample 1), 432 individuals for Tree 2 (sample 2), and 415 individuals for Tree 3 (random half of sample 1). Fourth, we specified that at each split, a node can be divided into only two nodes. A reason for developing binary trees is to mitigate selection bias towards categorical variables, especially those with more risk classes (Kim & Loh, 2001; Loh, 2002; Loh & Shih, 1997; Shih & Tsai, 2004). Note that multiway splits can be achieved by several binary splits (Hastie et al., 2009). We allowed risk classes to be used multiple times across ancestor splits: For example, the age/gender variable could first be used to split the data in a group with children and another with all remaining individuals. In a next split, the age/gender risk classes identifying these remaining

[14]A detailed list of all tree parameters can be provided upon request from the first author.

individuals can be used again to split this group further into a group with individuals younger than 65 years and another with individuals older than 65 years.

To test the influence of these additional user-defined parameters on the identification of interactions, we estimated a tree with an alternative specification for each of these parameters, ceteris paribus to Tree 1. Estimation of these alternative trees showed that a different specification of a parameter results into identification of another set of interactions, implying that trees are not robust (Appendix S1).

2.3 | Risk-equalization models

The interactions identified by each of the five trees were used as additional risk adjusters in the form of a dummy, resulting in five extended RE-models (*RE-model 1-5*, M = number of risk classes = 260, 223, 237, 277, and 139, respectively). The predictive performance of these extended RE-models was compared with that of the Dutch RE-model of 2014 (*RE-model 0, M = 132*). The dependent variable in all RE-models was annualized total observed health care expenses. All RE-models included an intercept and were estimated by OLS, with a weight for the enrollment period (see footnote 3). The coefficients of all RE-models were estimated on total sample 2, which were used to predict individual expenses on total sample 3.

2.4 | Comparative model evaluation

All RE-models were evaluated for all individuals in sample 3 and for selective groups based on an external dataset. At the sample level, the "adjusted R-squared" (R^2), "Cummings Prediction Measure" (CPM), and "Mean Absolute Prediction Error" (MAPE) were calculated for each RE-model. These measures-of-fit examine how well the models on average predict expenses for the total sample. For the calculation and a thorough discussion of these measures-of-fit, see van Veen, van Kleef, van de Ven, & van Vliet, 2015.

The "Mean Prediction Error" (MPE; i.e., the mean of predicted expenses minus observed expenses) was calculated for 46 selective groups based on the survey data, using an approach similar to other studies (Stam, van Vliet, & van de Ven, 2010, van Kleef et al., 2014). In the survey data, questions like "How do you rate your health status?" and "Do you have one of the following diseases?" were used to define selective groups. Most groups were defined by "yes/no" questions. Appendix S2 describes the definition of groups based on more than one question and/or more answer categories. The groups comprised an overrepresentation of high-risk individuals, for example, chronically ill. A two-sided *T* test was applied to test whether the MPEs on the groups were statistically significantly different from zero. To perform this test, the MPE were corrected in such a way that the overall MPE for each model was zero on the survey sample: The MPE was multiplied by a factor equal to the average predicted expenses divided by average observed expenses.

The survey sample can be considered reasonably representative for the Dutch population with respect to the percentage of individuals with a PCG, DCG, MHC-group, and DME-group (Table 1). In the survey sample, average age is lower than average age in the population, which may be the result of excluding nursing homes. Moreover, average observed expenses in the survey sample are €19 lower than average observed expenses in the population; however, this difference is not statistically significant.

3 | RESULTS

3.1 | Robustness of the identified interaction terms

Given the tree specifications and user-defined parameters, Tree 1 to Tree 5 identify 128, 105, 91, 145, and 7 interactions, respectively. Many higher order interactions with different levels are identified.[15] Tree 5 was restricted to third-order terms, while 85%–97% of the interactions identified by the other trees consist of higher-than-third-order terms, with a maximum of an 11th-order term by Tree 3 and Tree 4. In total, 2.3%, 3.1%, 1.6%, and 1.6% of the interactions of Tree 1 were identical to those of Trees 2, 3, 4, and 5, respectively, with no identical interactions across all trees. Consequently, it is not useful to thoroughly discuss which interactions are identified. Some risk classes that were generally used across trees are men ≥90 years, the MHC-group for no multiple-year high costs or for three consecutive years in the top 1.5%, the SES-classes for age ≥ 65 years, and a PCG for no use of drugs or for use of drugs for heart diseases, Crohn's disease, HIV/AIDS, transplantations, or brain/spinal cord diseases. Note that using these risk classes for defining a split implies defining the complementary group as well. These findings show that the trees are not robust in terms of the identification of interactions.

[15]The definition of all interactions of all trees can be provided upon request from the first author.

TABLE 3 The predictive performance of the Dutch RE-model 2014 and the RE-models extended with interaction terms at the sample level, on the total sample 3 of the administrative dataset (N = ~4.1 million)

Estimated RE-models	Adj.-R^2 (%)[a]	CPM (%)[b]	MAPE (Euros)[c]
RE-model 0 (Dutch model 2014)	25.56	24.98	1,569
RE-model 1(RE-model 0 + all 128 interactions by Tree 1)	26.20	25.11	1,566
RE-model 2 (RE-model 0 + all 91 interactions by Tree 2)	26.92	25.27	1,563
RE-model 3 (RE-model 0 + all 105 interactions by Tree 3)	26.09	25.18	1,565
RE-model 4 (RE-model 0 + all 145 interactions by Tree 4)	27.34	25.42	1,560
RE-model 5 (RE-model 0 + all 7 interactions by Tree 5)	25.64	24.98	1,569

Note. All RE-models were estimated on the total sample 2 of the administrative dataset (N = ~4.2 million) and expenses were predicted on the total sample 3 (N = ~4.1 million).

[a]Adj.-R^2 = adjusted-R-squared. The adj.-R^2 is calculated as one minus ratio of the variance of residual expenses divided by variance of observed expenses, adjusted for the number of variables used in the model.

[b]CPM = Cummings' Prediction Measure. The CPM is calculated as one minus the ratio of the MAPE to the mean absolute difference between observed expenses and average observed expenses.

[c]MAPE = Mean Absolute Prediction Error. The MAPE is calculated as the absolute difference between predicted expenses and observed expenses. The MAPE is rounded to the nearest Euro.

3.2 | Models' predictive performance

3.2.1 | Sample level

Table 3 shows that interactions can improve models' predictive performance at the sample level. Including interactions in the Dutch RE-model of 2014 increases the R^2 of 25.56% by 0.08 to 1.78 percentage points and the CPM of 24.98% by 0 to 0.44 percentage points, depending on the specification of the tree. For each extended RE-model, the increase in R^2 in percentage points is larger than the increase in CPM, indicating that the interactions specifically identify groups of individuals with relatively high expenses. The MAPE of the extended RE-models vary from €1,566 (RE-model 1) to €1,560 (RE-model 4), implying that the MAPE remains the same or reduces somewhat when interactions are included. Of all extended RE-models, RE-model 4 shows the best performance at the sample level, and therefore, we examine to what extent this model predicts expenses for the selective groups in order to indicate financial incentives for risk selection.

3.2.2 | Selective groups based on external survey information

For all selective groups average, observed expenses are (far) above average observed expenses in the survey sample. This implies that they contain a high proportion of relatively high-risk individuals. Note that individuals can be classified to multiple groups.

Table 4 shows that RE-model 0 underpredicts expenses for 18 groups, for example, for persons who have limitations in hearing, a low score on physical health scales, or a poor general health status (Table 4). Extending this RE-model with interactions improves the prediction of expenses for eight groups but deteriorates the prediction for nine, and for one group, the prediction does not change. These results are not statistically significant, except for those who contacted a home nurse or who used durable medical equipment in the past year. For these two groups, interactions result into a modest improvement.

The MPE for some specific high-risk groups are still statistically significantly different from zero under RE-model 4, implying that even with interactions, the RE-model does not lead to adequate predictions for these groups. For example, these are individuals with a poor general health status or OECD limitations in hearing. This leads to the conclusion that interactions mitigate financial incentives for risk selection for some groups, but they cannot remove them, and they even increase for others. The other extended RE-models provide similar results.[16]

For the remaining 28 groups that are not presented in Table 4, the MPE did not statistically significantly differ from zero for RE-model 0, implying that interactions could not further improve models' predictive performance on these groups (Appendix S3).

The MPE for the groups under the other alternative RE-models (RE-models 1, 2, 3, and 5) are not presented in Table 4. The MPE did not statistically significantly differ across these RE-models. The MPE's are comparable to those of RE-model 4. This implies that different tree-model specifications for defining the interactions do not lead to different conclusions regarding the predictive performance for the defined groups.

[16]For ease of interpretability, we did not present the results for all RE-models at all 46 groups. These results can be provided upon request from the first author.

TABLE 4 The "Mean Prediction Error" for RE-model 0 and RE-model 4 on groups for which the MPE is statistically significantly different from zero for RE-model 0, using the health survey sample ($\underline{N}$ = 16,141)

Groups (based on health survey data from 2010)	Size (%)	Mean observed expenses in year t, in Euros	Mean prediction error in 2011 (mean of predicted expenses minus observed expenses), in Euros		Additional compensation by Model 4 versus Model 1 (MPE Model 4–MPE Model 0), in Euros
			Model 0	Model 4	
Health care utilization (all respondents)					
Contact with home nurse (care) in the past year	1.4	9,336	−1,343*	−1,164	+179
Use of durable medical equipment	7.2	5,094	−527*	−455	+72
General health status (all respondents)					
General health status is poor	19.0	4,279	−380***	−391***	−11
At least one long-term disease	31.6	3,480	−334***	−321***	+13
Functional disabilities (age ≥ 12 years)					
OECD limitations in seeing	6.1	3,883	−712**	−694**	+18
OECD limitations in hearing	2.9	5,133	−1,207**	−1,232**	−25
Scores on SF-12 (age ≥ 12 years)					
A low score on physical health scales	19.0	4,476	−671***	−674***	−3
The lowest score on physical health scales	9.5	5,707	−835**	−859**	−24
Self-reported disease in the past year (age ≥ 12 years)					
Serious/persistent back problems or pain	10.6	3,488	−368**	−358*	+10
Serious bowel disorders, longer than 3 months	3.4	4,412	−745*	−705*	+30
Comorbidity (age ≥ 12 years)					
Three or more self-reported diseases or (chronic) disorder	17.5	4,212	−336**	−361**	−25
Health care utilization (all respondents)					
Hospitalization in the past year	6.6	5,773	−576**	−596**	−20
Prescribed drugs use in the past 14 days	35.7	3,133	−188***	−171**	+17
Contact general practitioner in the past year	72.0	2,068	−81*	−84**	−3
Contact medical specialist in the past year	37.9	3,114	−326***	−326***	0
Contact physiotherapist in the past year	21.8	2,934	−327***	−328***	−1
Hearing aid	3.4	4,760	−613*	−614*	−1
Limitations in daily activities (ADL) (age ≥ 55 years)					
At least one bad score on ADL scales	3.6	7,227	−640*	−638*	+2

***Statistically significantly different from zero with $p \leq .01$;

**Statistically significantly different from zero with $p \leq .05$;

*Statistically significantly different from zero with $p \leq .10$ (based on a two-sided T test).

Note. Predicted expenses in the survey sample were corrected in such a way that average MPE (predicted expenses minus observed expenses) on the total survey sample is zero for each RE-model. This was done to test the statistical significance of the MPEs from zero. Average predicted expenses for each model slightly deviated from average observed expenses in the evaluation sample, because the models were estimated on another sample.The last column presents how many additional money is distributed to this group as a result of including interactions in the Dutch RE-model of 2014. A positive value means that insurers receive a higher compensation for this group under Model 4 than under Model 0. This means that the extended model better predicts average expenses for this group and so, the under-compensation is closer to zero. A negative value implies that insurers receive a lower compensation for this group under Model 4 than under Model 0, implying that the under-compensation for this group increases under Model 4. The statistical significance for the difference in MPE between these two models is not of particular interest, because it is of interest whether the MPE under Model 4 is statistically significantly different from zero, see the fifth column for these results.

4 | CONCLUSIONS AND DISCUSSION

This study explored the predictive power of interaction terms ("interactions") between the risk classes in the Dutch risk-equalization (RE) model of 2014. The interactions were identified by regression trees in a first step and then were used as additional risk adjusters in this RE-model in a second step.

A first important finding is that interactions can improve the predictive performance of this RE-model overall and for specific groups in the population. This is consistent with findings by Buchner and colleagues (Buchner et al., 2015). In our study, inclusion of interactions in the RE-model increases the R^2-value of 25.56% by 0.08 to 1.78 percentage points and the CPM-value of 24.98% by 0 to 0.44 percentage points, and the MAPE of €1,569 decreases by €0 to €9, depending on the specification of the regression tree. This leads to the conclusion that interactions can mitigate financial incentives for risk selection. If the groups that are explicitly identified by the interactions are vulnerable to risk selection, this may be a good reason to include interactions in the RE model.

A second important finding is that including interactions in a sophisticated RE-model also deteriorates the prediction of expenses for some selective groups, which increases financial incentives for risk selection on these groups. Consequently, extending an existing RE-model with interactions is not enough to obtain an adequate RE-model. Further research is needed to investigate new types of risk adjusters and/ or methods to improve RE-models.

Consistent with the literature (Gail et al., 2009; Hastie et al., 2009; Strobl et al., 2009), the regression trees are not robust with respect to the identification of interactions. A different set of interactions is identified when other model specifications are used. Consequently, we cannot draw conclusions about which interactions should be used in practice. This study also shows that a different tree-model specification does not lead to statistically significantly different results with respect to model's predictive performance for specific groups, given our datasets used and the defined groups. Further research is needed to investigate the robustness of regression trees in combination with the predictive performance of RE-models.

In order to implement interactions in practice, other criteria play a role in addition to the predictive power, such as the right incentive structure for risk selection and efficiency or the opinion of medical experts, for example, interactions reflecting comorbidity. In practice, regression trees can be used differently. A first approach is to use them to identify complex, unknown associations in the dataset guided by statistical criteria, as done in this study. In a second step, the list of interactions identified are judged by medical experts and/ or policymakers in order to decide which interactions should be implemented. An advantage of this approach is that all interactions that are statistically relevant are identified, indicating the maximum potential for improving model's performance. A disadvantage of this approach is that the interactions identified may represent unexpected relationships that do not satisfy medical experts' or policymakers' criteria. An alternative approach is to incorporate the opinion of medical experts and/ or policymakers in the first step to specify the tree parameters. An advantage is that the list of interactions may adhere to the criteria that are important to them. A disadvantage is that this approach may lead to a limited set of interactions because the specified tree parameters limit the possibilities of finding statistically relevant interactions. Furthermore, unknown relationships that may be of interest could be overlooked.

The risk adjusters in the Dutch RE-model of 2014 are taken as given to identify the interactions. Risk adjusters in practice are chosen by statistical criteria as well as political and practical criteria. This may have limited the possibilities of identifying statistically relevant interactions but an advantage is that our interactions could be implemented immediately. For further research it is useful to explore finer categories of these risk adjusters; e.g. by using the detailed information underlying the pharmaceutical cost groups (PCGs) or diagnostic cost groups (DCGs). This detailed information may lead to interactions reflecting specific diagnoses of multi morbidities, in particular when medical experts are involved during the process of developing the interactions.

ACKNOWLEDGEMENTS

We gratefully acknowledge the Dutch Ministry of Health, Welfare, and Sports and the Dutch association for health insurers, “Zorgverzekeraars Nederland,” for their permission to use the administrative dataset for this research project. We also thank “Statistics Netherlands,” “Centraal Bureau voor de Statistiek,” for their permission to use the health survey data and their assistance with merging the administrative data with the health survey data. For comments on an earlier version of the paper, we thank Florian Buchner, Sonja Schillo, Jürgen Wasem, the editor, and two anonymous reviewers. The views presented in this study do not necessarily reflect those of the aforementioned organizations, entities, or persons.

CONFLICT OF INTEREST

No conflict of interests.

REFERENCES

Adams, E. K., Bronstein, J. M., & Raskind-Hood, C. (2002). Adjusted clinical groups: Predictive accuracy for Medicaid enrollees in three states. *Health Care Financing Review*, *24*(1), 43–61.

Ash, A., Porell, F., Gruenberg, L., Sawitz, E., & Beiser, A. (1989). Adjusting Medicare capitation payments using prior hospitalization data. *Health Care Financing Review*, *10*(4), 17–29.

Behrend, C., Buchner, F., Happich, M., Holle, R., Reitmeir, P., & Wasem, J. (2007). Risk-adjusted capitation payments: How well do principal inpatient diagnosis-based models work in the German situation? Results from a large data set. *European Journal of Health Economics*, *8*(1), 31–39.

Berk, R. A. (2006). An introduction to ensemble methods for data analysis. *Sociological Methods & Research*, *34*, 263–295.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. Monterey, Calif., U.S.A.: Wadsworth, Inc.

Buchner, F., Wasem, J., & Schillo, S. (2015). Regression trees identify relevant interactions: Can this improve the predictive performance of risk adjustment? *Health Economics*. https://doi.org/10.1002/hec.3277

Ellis, R. P., & Mookim, P. G. (2009). Cross-validation methods for risk adjustment models. Boston University working paper: July 1.

Fishman, P. A., Goodman, M. J., Hornbrook, M. C., Meenan, R. T., Bachman, D. J., & O'Keeffe Rosetti, M. C. (2003). Risk adjustment using automated ambulatory pharmacy data. The RxRisk model. *Medical Care*, *41*(1), 84–99.

Gail, M., Krickeberg, K., Samet, J., Tsiatis, A., & Wong, W. (2009). *Statistics for biology and health*. New York: Springer Science + Business Media, LLC.

Hastie, T. R., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer.

Hughes, J. S., Averill, R. F., Eisenhandler, J., Goldfield, N. I., Muldoon, J., Neff, J. M., & Gay, J. C. (2004). Clinical risk groups (CRGs): a classification system for risk-adjusted capitation-based payment and health care management. *Medical Care*, *42*(1), 81–90.

Kautter, J., Pope, G. C., & Keenan, P. (2014). Affordable care act risk adjustment: Overview, context, and challenges. *Medicare & Medicaid Research Review*, *4*(3), E1–E11.

Kim, H., & Loh, W. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, *96*(454), 589–604.

Kleef, R. C., van Vliet, R. C. J. A., & van de Ven, W. P. M. M. (2014). Risicoverevening 2014 voor somatische zorg. Analyse van uitkomsten op subgroepniveau. *Onderzoeksrapport iBMG*, Instituut Beleid & Management Gezondheidszorg, Erasmus Universiteit Rotterdam, Rotterdam.

Kronick, R., Gilmer, T., Dreyfus, T., & Lee, L. (2000). Improving health-based payment for Medicaid beneficiaries: CDPS. *Health Care Financing Review*, *21*(3), 29–64.

Lamers, L. M., & van Vliet, R. C. J. A. (2003). Health-based risk adjustment: Improving the pharmacy-based cost group model to reduce gaming possibilities. *European Journal of Health Economics*, *4*(2), 107–114.

Last, M., Maimon, O., & Minkov, E. (2002). Improving stability of decision trees. *International Journal of Pattern Recognition and Artificial Intelligence*, *16*, 2.

Loh, H. -W. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, *12*, 361–386.

Loh, H. -W., & Shih, Y. -S. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*, 815–840.

Newhouse, J. P. (1996). Reimbursing health plans and health providers: Efficiency in production versus selection. *Journal of Economic Literature XXXIV*, 1236–1263.

Oates, T., & Jensen, D. (1997). The effects of training set size on decision tree complexity. In Machine learning: Proceedings of the fourteenth international conference. *Morgan Kaufmann* (pp. 254–262).

Payne, S. M. C., Cebul, R. D., Singer, M. E., Krishnaswamy, J., & Gharrity, K. (2000). Comparison of risk-adjustment systems for the Medicaid-eligible disabled population. *Medical Care*, *38*(4), 422–432.

Pope, G. C., Ellis, R. P., Ash, A. S., Ayanian, J. Z., Bates, D. W., Burstin, H., … Wu, B. (2000). Diagnostic cost group hierarchical condition category models for Medicare risk adjustment. Final report to the Health Care Financing Administration under Contract No. 500-95-048. Waltham, MA.: Health Economics Research, Inc.; December 2000.

Pope, G. C., Ellis, R. P., Ash, A. S., Liu, C. -F., Ayanian, J. Z., Bates, D. W., … Ingber, M. J. (2000). Principal inpatient diagnostic cost group model for Medicare risk adjustment. *Health Care Financing Review*, *21*(3), 93–118.

Pope, G. C., Kautter, J., Ellis, R. P., Ash, A. S., Ayanian, J. Z., Iezzoni, L. I., … Robst, J. (2004). Risk adjustment of Medicare capitation payments using the CMS-HCC model. *Health Care Financing Review*, *25*(4), 119–141.

Robinson, J. W. (2008). Regression tree boosting to adjust health care cost predictions for diagnostic mix. *Health Research and Educational Trust*, *43*(2), 755–772.

Sarma, K. S. (2007). *Predictive modeling with SAS® Enterprise Miner™: Practical solutions for business applications*. Cary, NC: SAS Institute Inc.

Shih, Y. -S., & Tsai, H. -W. (2004). Variable selection bias in regression trees with constant fits. *Computational Statistics and Data Analysis*, *45*, 595–607.

Stam, P. J., van Vliet, R. C. J. A., & van de Ven, W. P. M. M. (2010). Diagnostic, pharmacy-based and self-reported health measures in risk equalization models. *Medical Care*, *48*(5), 448–457.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348.

van de Ven, W. P. M. M., & Ellis, R. P. (2000). Risk adjustment in competitive health plan markets. In A. Cutler, & J. P. Newhouse (Eds.), *Handbook of health economics* (First ed.) (pp. 755–845). B.V., Amsterdam: Elsevier Science.

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

Ven, W. P. M. M., van de Beck, K., van de Voorde, C., Wasem, J., & Zmora, I. (2007). Risk adjustment and risk selection in Europe: 6 years later. *Health Policy*, *83*, 162–179.

Zhao, Y., Ash, A. S., Ellis, R. P., Ayanian, J. Z., Pope, G. C., Bowen, B., & Weykuer, L. (2005). Predicting pharmacy costs and other medical costs using diagnoses and drug claims. *Medical Care*, *43*(1), 34–43.

Zhao, Y., Ellis, R. E., Ash, A. S., Calabrese, D., Ayanian, J. Z., Slaughter, J. P., … Bowen, B. (2001). Measuring population health risks using inpatient diagnoses and outpatient pharmacy data. *Health Services Research*, *36*(6), 180–193.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: van Veen SHCM, van Kleef RC, van de Ven WPMM, van Vliet RCJA. Exploring the predictive power of interaction terms in a sophisticated risk equalization model using regression trees. *Health Economics*. 2017;1–12. https://doi.org/10.1002/hec.3523