Eur J Health Econ (2015) 16:201–218
DOI 10.1007/s10198-014-0567-7

ORIGINAL PAPER

Improving the prediction model used in risk equalization: cost and diagnostic information from multiple prior years

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

Received: 8 October 2012 / Accepted: 14 January 2014 / Published online: 12 February 2014

Abstract Currently-used risk-equalization models do not adequately compensate insurers for predictable differences in individuals' health care expenses. Consequently, insurers face incentives for risk rating and risk selection, both of which jeopardize affordability of coverage, accessibility to health care, and quality of care. This study explores to what extent the predictive performance of the prediction model used in risk equalization can be improved by using additional administrative information on costs and diagnoses from three prior years. We analyze data from 13.8 million individuals in the Netherlands in the period 2006–2009. First, we show that there is potential for improving models' predictive performance at both the population and subgroup level by extending them with risk adjusters based on cost and/or diagnostic information from multiple prior years. Second, we show that even these extended models do not adequately compensate insurers. By using these extended models incentives for risk rating and risk selection can be reduced substantially but not removed completely. The extent to which risk-equalization models can be improved in practice may differ across countries, depending on the availability of data, the method chosen to calculate risk-adjusted payments, the value judgment by the regulator about risk factors for which the model should and should not compensate insurers, and the trade-off between risk selection and efficiency.

Keywords Competitive health care schemes · Health insurance · Risk equalization · Predictive performance

JEL Classification I13 · I18

S. H. C. M. van Veen (✉) · R. C. van Kleef · W. P. M. M. van de Ven · R. C. J. A. van Vliet
Institute Health Policy and Management, Erasmus University Rotterdam, PO box 1738, 3000 DR Rotterdam, The Netherlands
e-mail: vanveen@bmg.eur.nl

Introduction

Background

Several countries worldwide have implemented risk equalization (RE) into their (competitive) health insurance schemes. RE is a system of prospective risk-adjusted payments to compensate health insurers or health plans for predictable differences in individuals' health care expenses. The principal goals of RE are (1) to achieve affordability of health insurance for high-risk individuals and (2) to mitigate financial incentives for insurers to engage in risk selection [51]. The latter is particularly relevant for competitive health insurance schemes with premium regulation as found in Belgium, Germany, Israel, the Netherlands, and Switzerland.

Schokkaert and van de Voorde [36–38] have advocated that the calculation of risk-adjusted payments involves two steps. The first step focuses purely on the estimation of the prediction model, with the aim of explaining variation in individual health care expenses and to obtain accurate predictions, as far as is possible. Schokkaert and van de Voorde propose to include all relevant risk factors in the model, independent of whether the regulator desires compensation for those risk factors, in order to avoid (omitted-variables) bias in the predictions of individual expenses [36–38]. In the second step, the estimated model is used to calculate risk-adjusted payments. This step involves normative choices by the regulator on the appropriateness of incentives for efficiency and risk selection and on risk factors for which insurers should be compensated. If a regulator does not desire compensation for a risk factor, the effects of this risk factor can be neutralized in the calculation of the risk-adjusted payments; e.g., by using the average value of this factor or any other value identical for

all individuals in the population [36–38]. These normative choices on appropriateness of incentives and on risk factors for which insurers should and should not be compensated may be decided differently in different countries. The empirical analysis of our study focuses purely on the first step of the calculation of risk-adjusted payments; i.e., on the estimation of the prediction model.

Over the past two decades, the predictive performance of the models used in RE has improved substantially as a result of the development of diagnostic-based and pharmacy-based risk adjusters [1, 12, 13, 16, 17, 19, 22–24, 32, 33, 35], with over the past 5 years an increasing attention in the RE literature on the development of indicators for health status based on prior utilization or costs [e.g., 46, 47], and risk adjusters based on self-reported health or chronic conditions [e.g., 9, 13, 43]. Examples of diagnostic-based and pharmacy-based models are those used in Belgium, Germany, the Netherlands, and the US (Medicare). Several studies, however, have shown that even these sophisticated models do not adequately predict individual expenses, especially for high-risk individuals [4, 6, 48, 49]. Consequently, insurers receive risk-adjusted payments that are predictably too low for high-risk individuals and too high for low-risk individuals, which confronts insurers with incentives for risk rating and/or risk selection. Risk rating and risk selection both jeopardize affordability of coverage, accessibility to health care, and quality of health care [51, 52]. For example, insurers can select risks by offering less attractive benefits, or not contracting high-quality care, or providing poor services to high-risk subgroups [30, 51]. To mitigate incentives for risk rating and/or risk selection and to stimulate efficiency, further improvement of currently-used prediction models in RE is important.

Study objective and its contribution

This study endeavors to improve the prediction models used in RE by extending them with risk adjusters based on administrative information on costs and diagnoses from *multiple* prior years. Most of the currently used models use administrative data from one year to predict expenses in the next year. In 2012, the Dutch model has been extended with a risk adjuster for 'multiple-year high costs' [47, 49]. The Dutch model also includes risk adjusters based on diagnoses from previous year's hospitalizations, the so-called diagnostic cost groups (DCGs), and on previous year's use of prescribed drugs, the so-called pharmaceutical cost groups (PCGs). As studies have shown, the addition of risk adjusters based on costs and diagnostic information from multiple prior years may lead to more accurate predictions for individuals with systematically high expenses, such as the chronically ill [20, 21, 42, 47, 49]. Since most of the currently used models use 'only' information from one prior year and the Dutch model of 2012 uses in addition 'only' information on *total* prior costs (and not diagnoses from multiple prior years), it is expected that inclusion of additional risk adjusters using such information from multiple prior years could further improve models' predictive performance.

The present study makes two important contributions to the RE literature. First, this study develops two models: one that uses *diagnostic* information from multiple prior years and another that in addition uses *cost* information from multiple prior years. Comparing the predictive performance of these models with those of several (proxies for) currently used models will indicate the extent to which these models could potentially be improved by using administrative information on diagnoses and costs from multiple prior years. Second, assessing the predictive performance of these two newly developed models will indicate to what extent these models adjust payments for differences in individuals' expenses and so, whether these models would adequately compensate insurers.

This study uses an innovative approach. We used a very large administrative dataset covering almost the entire Dutch population (13.8 million observations) with lots of potentially relevant variables over multiple years. Using this dataset, we constructed a large array of multiyear cost-based and diagnostic-based adjusters, which have been used to develop two models. To specify the model using both cost-based and diagnostic-based adjusters, we used several variable-selection methods to select variables that contribute statistically significantly to models' predictive power. All models estimated in this study are evaluated on an external dataset with health survey information.

Our empirical analysis is limited to estimating prediction models used in RE and assessing the predictive performance of these models. This analysis does not focus on normative choices involved with the calculation of the risk-adjusted payments in practice, nor does it focus on other qualitative criteria used for deciding on the design of the model used in practice, such as feasibility in terms of necessary data, redistributional effects, or vulnerability to manipulation [51]. This implies that we estimate several prediction models and examine the fit between predicted expenses and observed expenses. The closer predicted expenses are to observed expenses, the better the model adjusts for the differences in individuals' observed expenses. It should be noted, however, that in practice a model with a better fit between predicted and observed expenses may not always be preferred over a model with a lower fit, because the payments to insurers or health plans do not have to (and cannot) adjust for all variation in individuals' observed expenses. There is a considerable amount of variation in observed expenses due to acute events (i.e., random variation), which is unpredictable and

for which insurers or health plans should not be compensated. In addition, there is variation in observed expenses due to risk factors for which the regulator desires compensation; the so-called compensation-type (C-type) risk factors (e.g., age, gender, need of health care related to health status), and risk factors for which compensation may not be desired; the so-called responsibility-type (R-type) risk factors (e.g., practice variation, inefficiency in provision of care, or moral hazard). Using information on costs and diagnoses from multiple prior years has been often debated in the RE literature and it has been applied in practice in only a (very) limited way for calculating risk-adjusted payments, because risk adjusters based on prior costs and/or prior utilization may reduce incentives for efficiency [e.g., 20, 21, 53, 54]. Following the approach of Schokkaert and van de Voorde [36, 37], we do not have to be concerned with these normative choices about C- and R-type risk factors in our empirical analysis, because we focus purely on improving the prediction model. Based on the models developed in this study, the regulator could decide which risk factors in the model are C- or R-type factors and then neutralize the effects of R-type risk factors in order to derive the risk-adjusted payments used in practice.

This study is relevant for all regulators and policy-makers in countries with a RE scheme or for those who want to incorporate RE in their health insurance scheme. Although this study uses administrative data from the Netherlands, regulators and policy-makers from other countries could learn from the findings of this study, because several models that are similar to currently used RE models have been evaluated. For this reason, the results of this study and the policy and methodological implications may be relevant for (most) countries with RE or those who are planning to implement RE. This study aims to indicate areas in which currently used prediction models in RE could be further improved.

The remainder of this article is structured as follows. First we describe the data and methods used in the empirical analysis, and then we present the results. Finally, we conclude and discuss these results, highlighting limitations of the study method, formulating points for further research, and addressing health-policy implications for regulators in countries with a RE scheme and for those who are planning to implement RE in their health insurance schemes.

Data and methods

Administrative data and health survey data

Two datasets were used for the empirical analysis. The first dataset contained individual-level administrative data for the Dutch population for the period 2006–2009. The sample analyzed in this study consisted of individuals who were enrolled, for a part or a full year, in each of the 4 years[1] ($N = 13.8$ million). For those individuals, we had the following three types of information for each year: (1) demographic information, including age, gender, region, source of income, and socio-economic status; (2) diagnostic information, including DCGs and PCGs, based on prior hospitalization and prior use of prescribed drugs respectively; and (3) cost information for several types of care. Total expenses are the sum of expenses on these different types of expenses. The administrative dataset is used for predicting individual expenses. The dependent variable in each of the estimated models is annual total health care expenses in the year 2009, which we refer to as prediction year t. Total expenses in year t were annualized and weighted by the fraction of the year the individual was enrolled.[2] For example, an individual who died after 3 months in year t and had 100 Euro expenses was given a weight of 0.25 and 400 Euro annual expenses. By applying this method, mean predicted expenses in year t equals mean observed expenses in year t. Table 1 shows some descriptive statistics. Mean total expenses in year t, $t-1$, $t-2$, and $t-3$ were 1,689 Euro, 1,639 Euro, 1,495 Euro, and 1,383 Euro, respectively. In the study population in year t, the average age was 41.5 years, 2.8 % of the individuals were classified into a DCG, and 17.7 % into a PCG, with 3.5 % having more than one PCG. In the Netherlands, individuals can be classified into only one DCG per year—the one with the highest follow-up costs—whereas individuals can be classified into more than one PCG in a year.

The second dataset contained information on self-reported health from year $t-1$ and is derived from a Dutch household survey, the *Permanent Survey of Living Conditions*. This survey is conducted each year on a representative sample of the Dutch population by "Statistics Netherlands".[3] It included detailed individual-level information on health status, household, and environment. We merged the administrative dataset with the survey data at

[1] Individuals who did not have continuous enrolment over the study period were excluded. Inclusion of deceased individuals is not useful for prediction purposes, but the exclusion of newborns may have moderately affected the generalizability of our results for the Dutch population.

[2] This weight is corrected for duplicate records in the dataset. Duplicate records were generated when merging the administrative data of 4 years due to switching behavior of individuals in prior years. Records of individuals who did not switch in year t, but who switched in 1 or more of the 3 years prior were copied (duplicates) when merging the administrative data of 4 years. These duplicate records were weighted by a value of 0.5 in the estimation of the model. There were no individuals who switched insurer more than once during 1 year (which would mean that more than two records would be generated during the merging process).

[3] "Statistics Netherlands" ("Centraal Bureau voor Statistiek") is an autonomous Dutch agency that collects and analyzes data.

Table 1 Mean of total observed expenses and some risk characteristics in year t and prior years, in the administrative data from the Dutch population of insured over a 4-year period ($N = 13.8$ million). *DCG* Diagnostic cost groups, *PCG* pharmaceutical cost groups

	Mean	SD
Total observed expenses (€)[a]		
Year t[b]	1,689	5,060
Year $t-1$	1,639	4,909
Year $t-2$	1,495	4,878
Year $t-3$	1,383	4,520
Risk characteristics		
Age (in years) in year t	41.5	22.24
Proportion male in year t	0.487	0.50
Proportion classified into a DCG[c] in year $t-1$	0.028	0.16
Proportion classified into a PCG in year $t-1$	0.177	0.38
Proportion classified into more than one PCG in year $t-1$	0.035	0.52

[a] The expenses in year t are annualized and weighted for the enrolment period. The expenses in the years $t-1$, $t-2$, and $t-3$ refer to observed expenses. All expenses are rounded to the nearest Euro

[b] The prediction year t is the year 2009. Years $t-1$, $t-2$, and $t-3$ are 2008, 2007, and 2006, respectively

[c] Individuals can be classified into only one DCG per year, the one with the highest follow-up costs

the individual level using an anonymous, unique identification variable ($N = 7{,}979$).[4] The health status information was used to define subgroups in the population to assess the predictive performance at the subgroup level. Given the administrative data and the health survey data, the following four-step procedure is applied to examine the additional value in terms of predictive performance when cost and diagnostic information from multiple prior years are used to predict expenses.

Model estimation

Model 1–4: proxies for currently used models

As a first step, four models were tested to compare the outcomes of the two newly developed models to the others. All independent variables in these models are dummy variables defining different risk classes in the population. Model 1 includes an intercept only in order to examine the situation where payments are not risk-adjusted but simply equal the mean expenses in year t. Model 2 includes variables for age interacted with gender (number of variables $= M = 39$). This demographic model can be considered as one of the simplest models used in practice. Model 3 includes the same risk adjusters as the Dutch model of 2011, which are age interacted with gender, region, source of income interacted with age, socio-economic status interacted with age, and DCGs and PCGs based on utilization in year $t-1$ ($M = 113$). "Appendix 1" describes the specification of these variables. A more detailed description is well-documented elsewhere [46]. Model 4 includes the same risk adjusters as the Dutch model of 2011; i.e., model 3, plus a risk adjuster for 'multiple-year high costs' defined over the 3 years prior ($M = 119$). Table 2 gives a description of the independent variables in each of the estimated models. It should be noted that the variables in these four models resulted from choices by the Dutch regulator on the C- and R-type risk factors, which does not hold for the two newly developed models.

Model 5: additional diagnostic information from three prior years

As a second step, we developed a model using diagnostic information from three prior years (Model 5). This model includes the same risk adjusters as model 3, extended with the DCGs and PCGs from year $t-2$ and $t-3$ ($M = 179$). The reference group in the model for the DCGs and PCGs in a certain year was the group of individuals without a DCG or a PCG, respectively, in that year.

Model 6: additional cost and diagnostic information from three prior years

As a third step, we developed a model using cost and diagnostic information from three prior years (Model 6). Using the administrative dataset, we defined 903 independent variables. We started with the same sets of variables as used in model 5; i.e., the set of variables included in model 3 ($M = 113$) plus the sets of dummy variables for DCGs and PCGs from year $t-2$ and $t-3$ ($M = 66$). Then, this model was extended with two sets of variables for prior costs. First, we defined dummy variables for percentiles of each type of expenses in year $t-1$, $t-2$, and $t-3$ ($M = 694$). We had information on the following types of expenses: hospital care, primary care, paramedical care, pharmaceuticals, durable medical equipment, transport in case of illness, dental care, obstetrical care, and maternity care. To define the percentiles, each type of expenses was divided into 20 risk classes, with each class representing 5 % of the population with positive expenses. The top 5 % of the distribution was further divided into five risk classes, with each class representing 1 % of the population with positive expenses. It is expected that these risk classes have strong predictive power, because being in the top 5 % of expenses in 1 year increases the likelihood of having high expenses in the next year(s) [15, 27]. All individuals with zero expenses per type of expenses were classified into a

[4] The administrative data is merged with the health survey data on the individual level according to Dutch privacy protection laws and regulations.

Table 2 Description of the independent variables for each estimated model. "Appendix 1" gives a more detailed description of the variables in models 1–4

Model	Description of the independent variables	Number of variables
Model 1 (no risk equalization)	A constant term (no independent variables)	0
Model 2 (demographic model)	39 dummy variables for age/gender risk classes	39
Model 3 (Dutch model of 2011)	Variables of model 2	113
	+9 dummy variables for region risk classes	
	+16 dummy variables for source of income/age risk classes	
	+11 dummy variables for socio-economic status/age risk classes	
	+13 dummy variables for DCGs from year $t-1$	
	+25 dummy variables for PCGs from year $t-1$	
Model 4 (Dutch model of 2012)	Variables of model 3	119
	+6 dummy variables for multi-year high costs risk classes	
Model 5 (multi-year health-based model)	Variables of model 3	179
	+26 dummy variables for DCGs in year $t-2$ and $t-3$[d]	
	+40 dummy variables for PCGs in year $t-2$ and $t-3$[e]	
Model 6 (multi-year health/cost-based model)	Variables of model 5	903[a]
	+694 dummy variables for percentiles of ten types of expenses from year $t-1$, $t-2$, and $t-3$[b,c]	
	+30 continuous variables for ten types of expenses from year $t-1$, $t-2$, and $t-3$	

[a] Not all 903 defined variables have significant predictive power and, therefore, are selected by the stepwise regression procedure to be used for predicting individual expenses. The stepwise regression procedure selected 562 variables ($\sim$ 62 %), using a 5 % significance level

[b] We had information on the following types of expenses: total expenses and expenses separately for hospital care, primary care, paramedical care, pharmaceuticals, durable medical equipment, transport in case of illness, dental care, obstetrical care, and maternity care

[c] For some type of expenses in a given year the threshold value of different percentiles was equivalent, which was due to insufficient variation in the left tail of the distribution of some type of expenses, e.g., expenses related to pharmaceuticals or durable medical equipment. When estimating the model, only one dummy variable for these equivalent percentiles was included. Therefore, the total number of defined variables differed across years. For years $t-1$, $t-2$, and $t-3$ we defined 225, 235, and 234 dummy variables, respectively, for percentiles of expenses

[d] For each year, we had 13 diagnostic cost groups

[e] For each year, we had 20 pharmaceutical cost groups

separate risk class, which was the reference group in the model for the set of dummy variables for percentiles per type of expenses. An individual was assigned to a risk class if the individual had expenses below or equal to the threshold value of the calculated percentile and higher than the threshold value of the previous percentile. Second, we added a set of continuous variables for each type of expenses in year $t-1$, $t-2$, and $t-3$ ($M = 30$). Dummy variables for percentiles of expenses as well as continuous variables were defined, because it was not known a priori which variables would have (more) predictive power.[5]

Stepwise regression methods were used to select only those variables with statistically significant predictive power. With 903 variables, not all of them may be relevant for predicting individual expenses. Stepwise regression methods are useful for selecting a subset of variables for purposes of prediction or exploratory data analysis [14, 31, 44]. Stepwise regression methods use a forward/backward selection procedure, which implies that variables can enter and leave at each step of the procedure, starting with the variable that yields the largest contribution to the model in terms of the F-statistic. At each step, the variable with the most significant F-statistic is added and any variable in the model producing a non-significant F-statistic is dropped. The procedure stops when no variable outside the model can make a significant partial contribution to the model and no variable in the model can be dropped without a significant loss in predictive power. We used a significance level of 0.05 to test the F-statistics.[6] In our analysis, we focused

[5] To examine to what extent percentiles of prior expenses and prior expenses continuous are 'substitutes', two other models were estimated; one model did not include percentiles for prior expenses and the other did not include continuous variables for prior expenses. These two models yielded adjusted R^2-values of 35.34 and 31.33 %, respectively. The adjusted R^2-value of model 6 is 35.98 %. These results indicate that continuous variables for expenses and dummy variables for percentiles of expenses both independently contribute to the predictive power of the model. Therefore, both types of variables were included in model 6.

[6] The described procedure is programmed in statistical software package SAS version 9.2.

primarily on *prediction* and not on hypothesis testing or causal interpretation to the effects of the independent variables. If the purpose were to draw statistical inferences about the effects of independent variables, the presence of (a high degree of) multicollinearity is of interest, because correlation among variables may influence the order of variable selection [14, 31]. For the purposes of prediction, however, multicollinearity is not of particular interest, because we are interested only in the predictive power of the model and not so much in which variables contribute (most) to the model.

A split-sample approach was applied in order to mitigate the influence of outlier observations and over-fitting of the data. The stepwise regression method selected a subset of variables that fit the data best. With this procedure, there is a risk of over-fitting the data when the same sample is used for both estimation of the model and prediction of expenses [3, 28]. Therefore, the total sample was split into a training sample and a validation sample. In the fourth step of the analysis, administrative data is merged with health survey data. To make maximum use of this data, we first assigned all respondents of the health survey to the validation sample, subsequently all other individuals were assigned randomly to either the training or validation sample, so that each sample contained approximately half of the total observations. This approach does not introduce selection bias and, therefore, both samples can be considered representative of the Dutch population that was enrolled during the study period.[7] All six models examined in this study were estimated on the training sample, and the coefficients of the variables in these models were used to predict individual expenses in the validation sample (model parameters of each estimated model can be provided on request to the first author).

All six models were assumed to be linear in the coefficients and included an intercept. The use of ordinary least squares (OLS)-models on untransformed data for predicting individual expenses has been discussed widely in literature, because OLS may not fit the distributional properties of health care expenses very well [5, 7, 10, 25, 26, 45]. We used an OLS-model on untransformed data to predict individual expenses for the following three reasons. First, OLS-models are easier to use and interpret than other models, such as two-part models (2PMs), generalized linear models (GLMs), or models based on (log-) transformed data. In the context of RE, this feature is highly important for regulators and policy-makers and therefore, OLS on untransformed data has been adopted widely in practice. Second, this study aims to examine the potential for improving currently used prediction models. To make a consistent comparison, we should estimate the models with the same estimation method as used in practice. Third, the analysis is based on a very large sample. Several studies have shown that when sample sizes are large (enough), OLS may provide the same model fit as more complicated models, such as 2PMs or GLMs [11, 18, 29, 34, 54]. Therefore, we expect that we would have found quite similar results with other estimation methods than OLS.

[7] Table 3 presents descriptive statistics of the training and validation-sample. Descriptive statistics of the total sample are not presented here but can be provided on request (contact the first author).

Model evaluation

As a fourth step, the predictive performance of the estimated models was assessed and compared at both the population and subgroup level. By doing so, it is possible to examine how well the models predict expenses for the total sample and for specific subgroups in the population of insured. At the population level, the adjusted *R*-squared (R^2) and mean absolute prediction error (MAPE) were calculated for each model. The MAPE was calculated as the average of the absolute differences between predicted expenses and observed expenses. Higher R^2-values and lower MAPE-values indicate a higher predictive performance of the model, since predicted expenses are closer to observed expenses.

Models' predictive performance at the subgroup level was assessed by the mean prediction error (MPE). The MPE was calculated as the average of the difference between predicted expenses and observed expenses, i.e., it is the average under- or over-prediction per individual in a subgroup. A model tends to perform better on subgroups defined by information from the training-sample than information from the validation sample and on subgroups matching (or highly correlated with) the risk cells of the model [8]. To perform a stronger test, we used an external dataset in the form of the health survey sample merged to the validation sample in order to evaluate models' predictive performance on subgroups ($N = 7{,}979$). The MPE on survey subgroups can provide a good indication of the extent to which models compensate insurers for differences in expenses between subgroups. This method is also applied in other studies [41, 42, 48, 49].

General demographic risk characteristics in the dataset used for the model evaluation at the subgroup level are comparable to those of the training- and validation-sample, providing evidence for the representativeness of the health survey respondents for the Dutch population (Table 3). However, there are three exceptions: the prevalence of young individuals with an age under 24, individuals with an age older than 25 but younger than 44 years, and individuals living at a home address with more than 15 persons. The first group is slightly overrepresented in the survey data while the second and third are underrepresented. The

Table 3 Descriptive statistics for individuals in the administrative data and the respondents of the health survey who matched successfully with the administrative data

General risk characteristics in year t[b]	Administrative data[a] Individuals in training-sample[c]	Administrative data[a] Individuals in validation-sample[d]	Survey data Respondents of health survey[e]
N (records)	6,999,827	7,001,379	8,091
N (individuals)	6,900,221	6,901,194	7,979
N (insured-years)[f]	6,855,800	6,856,876	7,938
Expenses			
Mean observed expenses	1,688	1,689	1,706
Mean predicted expenses of model 1	n.a.	1,689	1,689
Age/gender			
Men 0–24 years	13.80 %	13.81 %	15.67 %
Men 25–44 years	13.41 %	13.41 %	11.64 %
Men 45–64 years	14.10 %	14.11 %	13.57 %
Men 65–74 years	4.38 %	4.39 %	4.64 %
Men >75 years	2.96 %	2.96 %	2.95 %
Women 0–24 years	13.28 %	13.30 %	14.49 %
Women 25–44 years	13.78 %	13.73 %	12.68 %
Women 45–64 years	14.58 %	14.60 %	14.73 %
Women 65–74 years	4.78 %	4.77 %	5.00 %
Women >75 years	4.93 %	4.93 %	4.64 %
Region			
Cluster 1–5	50.19 %	50.18 %	47.55 %
Cluster 6–10	49.81 %	49.82 %	52.45 %
Source of income			
Individuals <18 years or >64 years	35.62 %	35.60 %	39.53 %
Disability benefit	5.36 %	5.36 %	4.85 %
Social security benefit	2.01 %	2.00 %	1.18 %
Self-employed	4.16 %	4.15 %	3.65 %
Others	52.85 %	52.88 %	50.80 %
Socio-economic status			
Living on a home address with ≥15 persons	1.41 %	1.39 %	0.38 %
Lowest income-class (deciles 1–3)	29.51 %	29.49 %	29.01 %
Middle income-class (deciles 4–7)	40.19 %	40.24 %	41.07 %
Highest income-class (deciles 8–10)	28.89 %	28.88 %	29.54 %
% classified in one or more PCGs	17.68 %	17.72 %	17.83 %
% classified in multiple PCGs	3.53 %	3.56 %	3.42 %
% classified in a DCG	2.81 %	2.82 %	2.64 %

[a] Individual-level administrative data from 2006 to 2009 is used

[b] Prediction year t is 2009

[c] The models are estimated on this sample

[d] Expenses of individuals are predicted on this sample

[e] Models' predictive performance at the subgroup level is assessed on this sample. The health survey is conducted in year $t-1$, 2008. The health survey dataset is merged with the administrative data (the validation-sample) on the individual level, using a unique, anonymous identification variable

[f] This is the sum of the weights for the fraction of the year the individual was enrolled. This number is lower than the number of individuals, because not all individuals have been insured for the full year

main reason for the latter is that the health survey is targeted mainly at individuals living in private households. Institutions, mental and nursing homes are excluded from the sample selection. Therefore, our results may not be representative for the subgroup of institutionalized individuals.[8]

Specifically, information on self-reported health status, (long-term) diseases and conditions, and health care utilization was used to construct 45 subgroups. These subgroups were defined in such a way that they include a relatively large proportion of high-risk individuals (e.g., chronically ill). These subgroups are comparable to those defined by van Kleef et al. [48, 49], Stam [40] and Stam and van de Ven [41]. The subgroups were identified by questions like: "How do you rate your health status?", "Do you have one of the following diseases?", "Do you have problems with performing a certain daily activity?". Most

[8] Based on an empirical analysis of Dutch administrative data from 2007, under-predictions varying from 300 Euro up to 1,400 Euro can be expected on subgroups with a relatively large proportion of institutionalized individuals [39].

subgroups were defined by 'yes/no'-questions. "Appendix 2" describes the definition of subgroups based on more than one question and/or more answer categories.

A (two-sided) t-test was applied to test whether the MPEs on subgroups are statistically significantly different from zero. To make this test relevant, the overall MPE for each model in the survey sample has to equal zero. This was, however, not the case; e.g., Table 3 shows that mean total observed expenses differs from mean total predicted expenses of model 1 in the survey sample. Therefore, the MPEs for each model in the survey sample were corrected as follows: individual observed expenses were raised by a factor equalling average predicted expenses in the survey sample divided by average observed expenses in the survey sample. These corrected MPEs were used to assess models' predictive performance on subgroups and to test the statistical significance of the MPEs.

Results

Predictive performance at the population level

The results in Table 4 show the predictive performance of the estimated models at the population level in terms of the adjusted R^2 and MAPE. These results show that the predictive performance of a model increases as more risk adjusters are added. Model 2 (i.e., a demographic model) has a R^2-value of 5.38 % and a MAPE of 1,808 Euro. As risk adjusters are added to model 2; i.e., socio-economic status interacted with age, source of income interacted with age, region, and DCGs and PCGs from one prior year, the R^2-value increases to 23.96 % and the MAPE-value reduces to 1,554 Euro. Adding risk adjusters for 'multiple-year high costs' to model 3 further increases the R^2-value to 28.54 % and the MAPE-value further reduces to 1,475 Euro. The R^2-value of model 5 is 24.84 % and the MAPE-value is 1,537 Euro, so that this model has a lower predictive performance than model 4. Based on this we may conclude that if model 3 is the benchmark and we aim to improve the predictive performance of the model, it may be more effective to include a risk adjuster based on cost information from multiple prior years than to include a risk adjuster based on diagnostic information from multiple prior years. When the model already uses a risk adjuster based on cost information from multiple prior years (model 4), its predictive performance could be further improved by approximately 8 percentage points in R^2-value by using additional cost and diagnostic information from three prior years. For models 1, 2, and 3 there is an even larger potential for improving the predictive performance by using cost and diagnostic information from multiple prior years. Consistent with other studies [2, 20, 54], these results confirm the predictive power of cost and diagnostic information from multiple prior years.

Table 4 Adjusted-R^2 and mean absolute prediction error (MAPE) of the estimated models

	Adjusted R^2 (in %)[b]	MAPE[c] (in Euro's)
Model 1 (no risk equalization)	0	1,997
Model 2 (demographic model)	5.38	1,808
Model 3 (Dutch model of 2011)	23.96	1,554
Model 4 (Dutch model of 2012)	28.54	1,475
Model 5 (multi-year health-based model)	24.84	1,537
Model 6 (multi-year health/cost-based model)	35.98	1,349

In this study, the adjusted R^2-value was equal to the (unadjusted) R^2-value, if rounded to two decimals. This is because the sample size is very large in comparison to the number of variables (=number of estimated parameters)

[a] The coefficients used for predicting individual expenses were obtained by estimating the models on the training-sample (random half of the dataset, approximately 7 million observations). The R^2-value was calculated on the validation-sample (complementary half of the dataset)

[b] Mean absolute prediction error (MAPE) was calculated as the average of the absolute differences between predicted expenses and observed expenses

Sensitivity analysis: specification model 6

To test the robustness of model 6, we performed a sensitivity analysis by changing the specification of the variable-selection procedure used for estimating this model. We estimated five alternative models. First, we re-estimated model 6 with two other variable-selection procedures than stepwise regression, namely backward elimination (alternative model 1) and forward selection (alternative model 2) [14, 44]. Second, we re-estimated model 6 with a significance level of 0.01 instead of 0.05 in order to examine whether the choice of significance level for entry and deletion of the variables influenced models' predictive performance (alternative model 3). Third, we re-estimated model 6 with the risk adjusters of model 3 as a starting point to which the stepwise regression method could add and delete variables based on cost and diagnostic information from three prior years; i.e., the risk adjusters of model 3 could not be deleted from the model. With this specification we examined whether it matters in terms of predictive performance if risk adjusters as used in practice are already included in the model. This procedure was applied twice, with one model using a significance level of 0.05 (alternative model 4) and the other using a level of 0.01 (alternative model 5). The predictive performance of these five alternative models appeared to be similar to those

of model 6 in terms R^2-values and MAPE-values; i.e., the R^2-values of the alternative models ranged from 35.976 to 35.978 %, with the R^2-value of model 6 being 35.976 % and the MAPE-value of the alternative models ranged from 1,348.87 Euro to 1,349.06 Euro, with the MAPE-value of model 6 being 1,348.96 Euro. These results indicate the robustness of the specification of model 6 as applied here for predicting individual expenses.

Predictive performance at the subgroup level

Based on analyzing the MPE-values of all models for the 45 subgroups, for 14 subgroups model 6 has reduced the MPE-value to such an extent that it is *not* statistically significantly different from zero, while all other models have produced statistically significant MPE-values, which means that adding cost and diagnostic information from three prior years has (statistically significantly) improved models' predictive performance (Table 5). For 7 subgroups all estimated models have produced statistically significant MPE-values, implying that adding risk adjusters based on cost and diagnostic information from three prior years is not sufficient to adequately predict expenses for these subgroups (Table 6). Finally, for 24 subgroups the MPE-value was not statistically significantly different from zero for one of the proxies for currently used models (models 1, 2, 3, or 4), implying that adding cost and diagnostic information from multiple prior years cannot further improve models' predictive performance statistically significantly ("Appendix 3"). In the remainder of this section, we focus purely on the first two types of results, i.e., on Tables 5 and 6.

For all defined subgroups expenses in year t are (far) above average expenses in the total sample in year t, indicating that all subgroups contain (as expected) a relatively high proportion of high-risk individuals. Further, for most subgroups the MPE has a negative value, which means that the models under-predict expenses for these subgroups. These under-predictions indicate that expenses for the complementary subgroups (i.e., the low-risk individuals) are over-predicted. Notice that positive MPE-values imply that the model over-predicts expenses for this subgroup. When interpreting the results in Tables 5 and 6, it should be taken into consideration that the same individual may occur in multiple subgroups.

The results in Table 5 show that models with more risk adjusters produce more accurate predictions at the subgroup level than models using less risk adjusters. For example, model 1 in Table 5 shows substantially negative MPE-values for all subgroups, all of them being statistically significantly different from zero. Compared to model 1, models 2, 3, and 4 further reduce the MPE-values for all subgroups, but statistically significant MPE-values still remain. Just as the performance at the population level, model 5 has a lower predictive performance than model 4. If model 3 is used as a benchmark, adding diagnostic information from three prior years improves the predictive performance for all subgroups: e.g., for individuals with OECD limitations in moving (age $\geq$12 years), individuals with a low score on the SF-12 scales (age $\geq$12 years), individuals with limitations in daily activities (age $\geq$55 years), or individuals who reported two or more diseases (age $\geq$12 - years). Model 4, however, further improves the performance for all subgroups in Table 5, which is due to the inclusion of a risk adjuster for 'multiple-year high costs'. Further, model 6 outperforms all other models on all subgroups in Table 5. The MPE-values on all subgroups in Table 5 have been reduced to such an extent that they are no longer statistically significantly different from zero. These results demonstrate that cost information from multiple prior years may be more effective in increasing models' predictive performance than diagnostic information from multiple prior years, given the dataset used in this study and the use of model 3 as the benchmark. Based on our results, we may conclude that using both cost and diagnostic information from multiple prior years may provide (statistically) significant improvements of models' predictive performance for several subgroups in the population.

However, the results in Table 6 show that model 6 (i.e. using cost and diagnostic information in addition to the Dutch model of 2012) still under-predicts expenses for several subgroups. Under-predictions (statistically significantly different from zero) remain for individuals who reported a poor general health status (age $\geq$12 years), one or more long-term diseases (age $\geq$12 years), a myocardial infarction or other serious heart disease (age $\geq$12 years), psoriasis (age $\geq$12 years), other long-term disease or disorder than migraine or other serious headaches, vascular constriction in stomach or legs, asthma or chronic bronchitis, chronic eczema, dizziness with falling down, or serious bowel disorders longer than 3 months (age $\geq$12 - years), three or more self-reported diseases or disorders (age $\geq$ 12 years), or use of complete dentures (age $\geq$16 - years). Apparently, these subgroups are not accurately identified by the additional risk adjusters based on costs and diagnoses from hospitalizations and use of prescribed drugs in three prior years.

Discussion

Methodological limitations and points for further research

The empirical analysis and the data used to illustrate the potential for improving the predictive performance of

Table 5 Subgroups for which the mean prediction error in year t is not statistically significantly different from zero for model 6. In this study, the prediction year t is 2009. The column of total expenses presents the corrected total expenses. Total expenses and predicted expenses in the sample with health survey information were corrected in such a way that the average MPE on the total survey sample is zero. This was done to test the statistical significance of the MPEs from zero. By doing so, the column with total expenses in year t minus the column with the MPEs of model 1 results into the same number for each group, namely total average expenses in year t (1,689 Euro)

Subgroups (based on health survey data from year $t-1$)	Mean prediction error in year t (=mean of [predicted expenses minus observed expenses]) (€)							
	Size, in %	Mean total expenses in year t, (in Euro's)	Model 1 (no risk equalization)	Model 2 (demographic model)	Model 3 (Dutch model of 2011)	Model 4 (Dutch model of 2012)	Model 5 (multi-year health-based model)	Model 6 (multiple year health/cost-based model)
Functional disabilities (age ≥12 years)								
OECD limitations in moving	6.6	5,743	−4,054***	−2,581***	−1,132***	−712*	−1,053**	−573
Scores on SF-12 (age ≥12 years)								
The lowest score on physical health scales	7.4	4,951	−3,262***	−2,283***	−1,121***	−727**	−989***	−426
A low score on physical health scales	14.6	3,943	−2,254***	−1,447***	−729***	−559***	−651***	−309
The lowest score on mental health scales	7.3	3,133	−1,444***	−1,151***	−692**	−668**	−642**	−465
A low score on mental health scales	14.6	2,585	−896***	−704***	−352*	−364**	−330*	−285
Limitations in daily activities (ADL) (age ≥55 years)								
At least one bad score on ADL scales	6.0	5,830	−4,141***	−2,478***	−1,093**	−504	−951**	−65
Self-reported disease or disorder, in the last year (age ≥12 years)								
Hypertension	12.6	3,709	−2,020***	−814***	−438**	−449**	−436**	−315
Urine incontinence	3.9	4,532	−2,843***	−1,640***	−1,199***	−865**	−1,226***	−618
Co-morbidity (age ≥12 years)								
Two self-reported diseases or (chronic) disorder	28.3	3,445	−1,756***	−994***	−476***	−417***	−430***	−206
Health care utilization (all respondents)								
Contact medical specialist in the past year	41.4	2,879	−1,190***	−852***	−474***	−369***	−452***	−112
Contact physiotherapist in the past year	19.8	2,549	−860***	−617***	−416***	−266**	−383***	−10
Prescribed drugs use in the past 14 days	40.6	3,050	−1,361***	−703***	−290***	−268***	−264***	−80
Health care utilization (age ≥4 years)								
Hearing-aid	3.6	5,017	−3,328***	−1,221**	−1,116**	−835*	−1,089**	−604
Health care utilization (age ≥12 years)								
Durable medical equipment	5.8	5,279	−3,590***	−2,292***	−1,451***	−966**	−1,374***	−603

*** Statistically significantly different from zero with $P \leq 0.01$; ** statistically significantly different from zero with $P \leq 0.05$; * statistically significantly different from zero with $P \leq 0.10$ (based on a two-sided t-test)

Table 6 Subgroups for which the mean prediction error in year t is statistically significantly different from zero for model 6. In this study, the prediction year t is 2009. The column of total expenses presents the corrected total expenses. Total expenses and predicted expenses in the sample with health survey information were corrected in such a way that the average MPE on the total survey sample is zero. This was done to test the statistical significance of the MPEs from zero. By doing so, the column with total expenses in year t minus the column with the MPEs of model 1 results into the same number for each group, namely total average expenses in year t (1,689 Euro)

Subgroups (based on health survey data from year $t-1$)	Mean prediction error in year t (=mean of [predicted expenses minus observed expenses]) (€)							
	Size, in %	Mean total expenses in year t, in Euro's	Model 1 (no risk equalization)	Model 2 (demographic model)	Model 3 (Dutch model of 2011)	Model 4 (Dutch model of 2012)	Model 5 (multi-year health-based model)	Model 6 (multiple year health/cost-based model)
General health status (all respondents)								
General health status is poor	20.2	4,207	−2,518***	−1,811***	−883***	−748***	−836***	−464**
At least one long-term disease	39.4	3,248	−1,559***	−1,098***	−512***	−425***	−476***	−221**
Presence of disease or disorder (age ≥12 years)								
Myocardial infarction or other serious heart disease (ever)	2.0	8,790	−7,101***	−5,298***	−3,386***	−3,068***	−3,143***	−2,649**
Self-reported disease or disorder, in the last year (age ≥12 years)								
Psoriasis	1.9	2,273	−584*	−155	65	357	107	429*
Other long-term disease or disorder	8.7	4,225	−2,536***	−1,991***	−970***	−750***	−889***	−489*
Co-morbidity (age ≥ 12 years)								
Three or more self-reported diseases or (chronic) disorder	16.4	4,388	−2,699***	−1,679***	−757***	−597***	−680***	−368*
Health care utilization (age ≥16 years)								
Complete dentures	13.4	4,289	−2,600***	−823***	−490**	−509**	−469*	−463*

*** Statistically significantly different from zero with $P \leq 0.01$; ** statistically significantly different from zero with $P \leq 0.05$; * statistically significantly different from zero with $P \leq 0.10$ (based on a two-sided t-test)

models in RE using cost and diagnostic information from multiple prior years have certain drawbacks. First of all, even though a large dataset is used, which is representative for the Dutch population, the dataset is restricted to a time period of three prior years. It is expected that cost and diagnostic information from more than three prior years could further improve models' predictive performance [15, 21, 27]. It is relevant to investigate how many years of lagged cost and diagnostic information would still have statistically significant predictive power in the estimation year. Such research may provide useful insights into the persistence of under-predicting expenses for certain high-risk groups in the population, which can indicate methods to further improve currently used prediction models in RE.

Second, our empirical analysis focused on improving models' predictive performance by using cost and diagnostic information from multiple prior years. However, other information not available in our dataset may also be useful for further improving the models, such as outpatient diagnostic information [50]. Our analysis is restricted in this sense and in practice there may be (many) more methods to further improve the prediction models. A relevant question is which other types of information than cost and diagnostic information from multiple prior years are available and how this information could be used to further improve the prediction models.

Third, the predictive performance of the model may depend on the statistical method chosen to predict individuals' expenses. We confined ourselves to the method used in practice, i.e., OLS, even though other statistical methods have been advocated in the literature [e.g., 5, 7, 10, 25, 26, 45]. To our knowledge, there is no empirical evidence on the predictive performance of transformed and/or nonlinear models based on millions of observations, compared to those of OLS models on untransformed data. Further research could provide pertinent evidence by investigating whether models' predictive performance can be further improved using a method other than those currently used in practice using large datasets (i.e., datasets with millions of observations). Moreover, further research is needed to investigate whether there is an additive or multiplicative relationship between risk adjusters based on cost and diagnostic information from multiple prior years. In this study, only additive relationships have been examined. Such research may result in further improvement of prediction models used in RE.

Health-policy implications

As Schokkaert and van de Voorde [36–38] have advocated, the calculation of risk-adjusted payments used in practice involves two steps. In the first step, the model is estimated with the aim of explaining variation in individual health care expenses and to obtain predictions that are as accurate as possible. The second step uses the estimated model to calculate risk-adjusted payments, which involves normative choices by the regulator on the appropriateness of incentives for risk selection and efficiency and on risk factors for which insurers should and should not be compensated. The empirical analysis of this study was restricted to the estimation of the prediction model. Consequently, we may not be able to draw definitive conclusions as to the extent to which currently used RE models can be improved in practice. Our findings should be interpreted bearing the following points in mind.

First, the extent to which currently used RE models can be improved may depend on the degree to which the risk adjusters satisfy the criteria of fairness, appropriateness of incentives for efficiency and selection, and feasibility. In our empirical analysis, we did not consider the fairness-criterion of the used risk adjusters in the two newly developed models: i.e., we did not distinguish risk factors for which the regulator desires compensation (C-type risk factors), and risk factors for which the regulator does not desire compensation (R-type risk factors) [37]. According to the approach of Schokkaert and van de Voorde [36–38], both C- and R-type risk factors should be included in the model in the first step of the calculation, instead of omitting these R-type risk factors, in order to avoid (omitted-variables) bias in the predictions. In the second step, the effects of these R-type risk factors can be neutralized, e.g., by using the average value of this risk factor or using the same value for all individuals in the population. Following this approach, regulators could use the models developed in this study by deciding which risk factors in the models are C- or R-type factors in order to neutralize the effects of R-type risk factors in the second step, and thus derive the risk-adjusted payments used in practice. Note that the choice of C-type and R-type risk factors involves a value judgment by regulators, which may be decided differently in different contexts by different regulators.

Note, however, that if regulators decide not to use cost and diagnostic information in the second step of the calculation of the risk-adjusted payments, because using this information may reduce incentives for efficiency, incentives for risk selection may increase compared to using this information in the calculation of the risk-adjusted payments. This trade-off between reducing incentives for risk selection and maintaining incentives for efficiency is inevitable as long as there are no better alternatives for risk adjusters than using cost and diagnostic information from multiple prior years. In the event that the regulator considers the incentives for risk selection to be too large compared to the reduced incentives for efficiency, information on costs and/or diagnoses from multiple prior years can be used in the second step of the calculation of the risk-

adjusted payments. In this case, restrictions could be placed on the risk adjusters based on prior costs and/or diagnoses in order to mitigate the reduction in incentives for efficiency. Examples are the thresholds on the 'Defined Daily Dose' for the PCGs and the requirement for the risk adjuster 'multiple-year high costs' that an individual is in the top 15 % for at least two of three consecutive years.

An advantage of the use of cost and diagnostic information from multiple prior years is that this type of information is, in most situations, already available in the administrative files of (Dutch) insurers or health plans. This means that it does not require a large additional administrative burden for collecting this information. In most situations, regulators and policy-makers could relatively easily improve the predictive performance of currently used models by including cost and diagnostic information from multiple prior years.

Conclusions

This study has explored the potential for improving the prediction models used in RE in competitive health insurance schemes. This study makes two important contributions. First, it shows that the predictive performance of currently used models can be improved by extending these models with risk adjusters based on cost and diagnostic information from multiple prior years. Compared to the Dutch model of 2012, the predictive performance of the model in terms of R^2-value could potentially be improved by 8 percentage points at the population level. At the subgroup level, models' predictive performance could also potentially be improved: e.g., improvements can be expected on groups of individuals who reported OECD limitations on moving, a low score on one of the SF-12 health scales, who have limitations in daily activities, or who have two or more diseases or (chronic) conditions. The second contribution of this study is that even a model using additional cost and diagnostic information from multiple prior years does not adjust for all differences in individuals' health care expenses, implying that there are still under-predictions (that are statistically significantly different from zero) for certain high-risk subgroups in the population: e.g., under-predictions remain for groups of individuals with a poor general health status, who have three or more diseases or (chronic) conditions, or who use complete dentures.

To conclude, currently used RE models do not adequately compensate insurers for predictable differences in individuals' health care expenses, which faces insurers with incentives for risk rating and risk selection, both of which jeopardize affordability of coverage, accessibility of health care, and quality of care. This study shows that these incentives for risk rating and risk selection could potentially be (substantially) reduced by further improving the predictive performance of the model using cost and diagnostic information from multiple prior years, but that even using this information does not remove these incentives completely. The extent to which currently used RE models can be improved in practice to the level of the two models developed in this study may differ across countries, depending on the availability of data, the method chosen to calculate risk-adjusted payments, the value judgment by the regulator about risk factors for which the model should and should not compensate insurers, and the trade-off between risk selection and efficiency.

Acknowledgments The authors gratefully acknowledge the Dutch Ministry of Health, Welfare and Sport and the national association of Dutch health insurers ("Zorgverzekeraars Nederland") for their permission to use administrative data for this study. In addition, we gratefully thank "Statistics Netherlands" ("Centraal Bureau voor Statistiek") for providing access to the health survey data. For their helpful comments on an earlier draft, we would gratefully thank the members of the Risk Adjustment Network and the two anonymous referees. The opinions in this article are those of the authors and do not necessarily reflect those of the above-mentioned organisations and individuals.

Appendix 1

See Table 7.

Table 7 Definition of risk adjusters included in estimated RE-models

Risk adjuster	Definition	Number of risk classes in the model[a]
Age/gender	40 risk classes (i.e., 20 risk classes for male and 20 risk classes for female), with age in 5-year classes, starting from 0 years, 1–4 years, 5–9 years, 10–14 year, 15–17 years, 18–24 years up to an age of 90. Individuals older than 90 years old are included in a separate risk class.	39
Region	10 risk classes, each class each class consists of a cluster—not necessarily adjacent—zip codes areas	9
Source of income/age	17 risk classes for source of income, with 4 categories of source of income (self-employment, disability benefits, unemployment benefits and social security benefits), interacted with 4 classes of age (15–34 years, 34–44 years, 45–54 years and 55–64 years). There is a separate risk class for individuals younger than 14 years or older than 64 years old	16

Table 7 continued

Risk adjuster	Definition	Number of risk classes in the model[a]
Socio-economic status/age	12 risk classes, with 4 socio-economic classes: SES 0 is for individuals living on 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 distribution, and SES 3 is for individuals in household with an income in the highest three deciles of the distribution, interacted with 3 age classes of 0–14 years, 15 to 64 years and individuals older than 65 years	11
PCG	26 risk classes. Individuals are assigned to a PCG when they used at least 180 daily dosages of a specific drug in the previous year. Individuals with no PCG were classified in PCG 0	25
DCG	14 risk classes. Individuals were assigned to a DCG when they had a hospital admission in the last year for a specific diagnosis. Individuals with no hospital admission were classified in DCG 0	13
Multi-year high costs	7 risk classes: three consecutive years in the top 15 %, top 10 %, top 7 %, top 4 %, top 1.5 % of total expenses, 2 years in top 15 % of total expenses and a separate class for those individuals who do not have high expenses in multiple years	6

[a] The number of variables included in the model is always one less than the number of defined risk classes, because one variable for each type of risk adjuster was a reference group for all included dummy variables per risk adjuster

Appendix 2

See Table 8.

Table 8 Description of all subgroups based on more than one question and/or more answer categories of the health survey

Subgroups	Definition
General health status	
A bad self-reported health status	The following question is answered with "bad or "very bad": "How do you rate your health status?"
Obesities	Obesities according to the Quetelet index, individuals with a BMI > 30
At least one long-term disease	The following question is answered with "yes": "Do you have one or more long-term disease?"
Functional disabilities	
OECD limitation in hearing	At least one of the following questions is answered with "yes, but with many difficulties" or "no, I can't": "Can you follow a conversation in a group of three or more persons?"; "Can you hold a conversation with another person?"
OECD limitation in seeing	At least one of the following questions is answered with "yes, but with many difficulties" or "no, I can't": "Can you read small letters in the newspaper?"; "Can you recognize someone at a distance of four meters?"
OECD limitation in moving	At least one of the following questions is answered with "yes, but with many difficulties" or "no, I can't": "Can you lift a weight of 5 kilo's?"; "When you are standing, can you bent down and lift something from the ground?"; "Can you walk for a distance of 400 meters uninterrupted?"
OECD limitation in talking	The following question is answered with "yes, but with many difficulties" or "no, I can't": "Can you speak intelligible?"
OECD limitation in eating	The following question is answered with "yes, but with many difficulties" or "no, I can't": "Can you bite and chew?"
Scores on SF-12	
The worst score/a bad score on physical health scales	Individuals with the worst or a bad score on the SF-12 physical component summary scale [55]
The worst score/a bad score on mental health scales	Individuals with the worst or a bad score on the SF-12 mental component summary scale [55]
Limitations in daily activities	
At least one bad score on ADL scales	At least one of the following questions is answered with "yes, but with many difficulties" or "no, I can't": "Can you eat and drink?"; "Can you come in and out of a chair?"; "Can you go to and come out bed?"; "Can you dress up and undress yourself?"; "Can you move inside your house?"; "Can you climb stairs?"; "Can you go in and out of your house?"; "Can you move outside your house?"; "Can you wash your hands and face?"; "Can you wash your body?"

Table 8 continued

Subgroups	Definition
Co-morbidity	
Two self-reported diseases or (chronic) disorders	Self-reported diseases or disorders on the questions: Do you have Diabetes Mellitus?, Did you have a stroke or brain infarction? Did you have a heart infarction or any other serious heart disease?, Did you have cancer?, Did you have migraine or serious headaches regularly in the last 12 months?, Did you have a high blood pressure in the last 12 months?, Did you have a narrowing of the blood vessels in your stomach or legs in the last 12 months?, Did you have asthma, bronchitis or lung emphysema in the last 12 months?, Did you have psoriasis in the last 12 months?, Did you have chronic eczema in the last 12 months?, Did you have a serious bowel disorder that persisted more than 3 months in the last 12 months?, Did you have involuntary urine loss in the last 12 months?, Did you have arthrosis of hips or knees in the last 12 months?, Do you have chronic arthrosis (rheumatoid arthritis)?, Did you have serious or persistent back problems or back pain in the last 12 months?, Did you have serious or persistent problems of neck or shoulder in the last 12 months?, Did you have serious or persistent problems of hand, wrist or elbow in the last 12 months?, Did you have another long-term disease or disorder?
Three or more self-reported disease or (chronic) disorders	
Health care utilization	
Durable medical equipment	At least one of the following questions is answered with "always": "Do you use an aid for walking (walker)?"; "Do you use a wheelchair (hand or electronic)?"; "Do you use an orthopaedic shoe?"; "Do you use a prosthesis (arm or leg)?"; "Do you use a splint?"; "How many times do you use things for urine incontinence?"; "How many times do you use a catheter?"; "How many times do you use a colostomy or things for urine or defecation?"

Appendix 3

See Table 9.

Table 9 Subgroups for which the mean prediction error in year t was already not statistically significantly different from zero for model 1, 2, 3, or 4. In this study, the prediction year t is 2009. The column of total expenses presents the corrected total expenses. Total expenses and predicted expenses in the sample with health survey information were corrected in such a way that the average MPE on the total survey sample is zero. This was done to test the statistical significance of the MPEs from zero. By doing so, the column with total expenses in year t minus the column with the MPEs of model 1 results into the same number for each group, namely total average expenses in year t (1,689 Euro)

Subgroups (based on health survey data from year $t-1$)	Mean prediction error in year t (=mean of [predicted expenses minus observed expenses]), in euro's							
	Size (%)	Mean total expenses in year t (in Euro's)	Model 1 (no risk equalization)	Model 2 (demographic model)	Model 3 (Dutch model of 2011)	Model 4 (Dutch model of 2012)	Model 5 (multi-year health-based model)	Model 6 (multiple year health/cost-based model)
General health status (all respondents)								
Obesity (age >30 BMI)	11.8	2,700	−1,011***	−581***	−179	−161	−175	−39
Functional disabilities (age ≥12 years)								
OECD limitations in hearing	2.5	3,606	−1917***	−590	−303	−168	−272	−55
OECD limitations in seeing	3.8	3,055	−1,366***	−483	71	213	68	230
OECD limitations in talking	0.2	2,099	−410	−396	−261	−241	−98	88
OECD limitations in eating	3.6	4,177	−2,488***	−1,056*	−445	−309	−338	−255
Presence of disease or disorder (age ≥12 years)								
Diabetes mellitus	4.1	4,757	−3,068***	−1,645***	59	203	156	357
Stroke, brain infarction (ever)	2.1	5,383	−3,694***	−1,878***	−997	−680	−905	−305
Some type of cancer (ever)	4.8	4,509	−2,820***	−1,364***	−681**	−403	−433	−205

Table 9 continued

Subgroups (based on health survey data from year $t-1$)	Mean prediction error in year t (=mean of [predicted expenses minus observed expenses]), in euro's							
	Size (%)	Mean total expenses in year t (in Euro's)	Model 1 (no risk equalization)	Model 2 (demographic model)	Model 3 (Dutch model of 2011)	Model 4 (Dutch model of 2012)	Model 5 (multi-year health-based model)	Model 6 (multiple year health/cost-based model)
Self-reported disease or disorder, in the last year (age ≥12 years)								
Migraine or serious headaches regularly	10.8	1,929	−240	−219	−145	−154	−124	−30
Vascular constriction (in stomach or legs)	1.9	5,769	−4,080***	−2,353**	−1,066	−778	−929	−469
Asthma, chronic bronchitis, lung emphysema	6.3	3,594	−1,905***	−1,376***	−468	−403	−368	−247
Chronic eczema	3.1	1,972	−283	−190	−174	−212	−186	−86
Dizziness with falling down	2.3	4,186	−2,497***	−1,515**	−490	−365	−469	−296
Serious bowel disorders, longer than 3 months	2.8	3,616	−1,927***	−1,402***	−677*	−509	−626*	−82
Self-reported disease or disorder, in the last year (age ≥12 years)								
Arthrosis of hips or knees	10.8	3,653	−1,964***	−665***	−284	−253	−263	−85
Rheumatoid arthritis	4.2	4,222	−2,533***	−1,521***	−603*	−550	−589	−325
Serious/persistent back problems or pain	8.6	2,795	−1,106***	−521**	−205	−155	−239	14
Serious/persistent problems of neck or shoulder	8.0	2,636	−947***	−364**	−151	−48	−137	118
Serious/persistent problems of hand, wrist or elbow	4.7	3,335	−1,646***	−974***	−367	−173	−375	7
Health care utilization (all respondents)								
Contact general practitioner in the past year	73.2	1,977	−288***	−174***	−95	−83	−92	4
Hospitalization in the past year	6.6	4,615	−2,926***	−2,288***	−917***	−406	−639*	128
Contact with visiting (home) nurse	1.3	7,284	−5,595***	−4,096***	−1,730*	−554	−1,382	208
Health care utilization (age ≥4 years)								
Glasses or contact lenses	38.9	2,403	−714***	−110	−73	−78	−78	−51
Health care utilization (age ≥14 years)								
Home help (assistance)	3.0	5,907	−4,218***	−2,124***	−831	64	−621	308

*** Statistically significantly different from zero with $P \leq 0.01$; ** statistically significantly different from zero with $P \leq 0.05$; * statistically significantly different from zero with $P \leq 0.10$ (based on a two-sided t-test)

References

1. Adams, E.K., Bronstein, J.M., Raskind-Hood, C.: Adjusted clinical groups: predictive accuracy for medicaid enrollees in three states. Health Care Financ. Rev. **24**, 43–61 (2002)
2. Ash, A., Porell, F., Gruenberg, L., Sawitz, E., Beiser, A.: Adjusting medicare capitation payments using prior hospitalization data. Health Care Financ. Rev. **10**, 17–29 (1989)
3. Babyak, M.A.: What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom. Med. **66**, 411–421 (2004)
4. Barry, C.L., Weiner, J.P., Lemke, K., Busch, S.H.: Risk adjustment in health insurance exchanges for individuals with mental illness. Am. J. Psychiatry **169**, 704–709 (2012)
5. Basu, A., Manning, W.G.: Issues for the next generation of health care costs analyses. Med. Care **47**, S109–S114 (2009)
6. Behrend, C., Buchner, F., Happich, M., Holle, R., Reitmeir, P., Wasem, J.: Risk-adjusted capitation payments: how well do principal inpatient diagnosis-based models work in the German situation? Results from a large data set. Eur. J. Health Econ. **8**, 31–39 (2007)
7. Buntin, M.B., Zaslavsky, A.M.: Too much abo about two-part models and transformation? Comparing methods of modeling medicare expenditures. J. Health Econ. **23**, 525–542 (2004)
8. Cumming, R.B., Knutson, D., Cameron, B.A., Derrick, B.: Claims-Based Methods of Health Risk Assessment for Commercial Populations. Research Report, Society of Actuaries, Milliman, Minneapolis (2002)
9. DeSlavo, K.B., Jones, T.M., Peabody, J., McDonald, J., Fihn, S., Fan, V., He, J., Muntner, P.: Health care expenditure prediction with a single item, self-rated health measure. Med. Care **47**(4), 440–447 (2009)
10. Duan, N., Manning, W.G., Morris, C.N., Newhouse, J.P.: A comparison of alternative models for the demand for medical care. J. Bus. Econ. Stat. **1**(2), 115–126 (1983)
11. Dunn, G., Mirandola, M., Amaddeo, F., Tansella, M.: Describing, explaining or predicting mental health care costs: a guide to regression models: methodological review. Brit. J. Psychiat. **183**, 398–404 (2003)
12. Fishman, P.A., Goodman, M.J., Hornbook, M.C., Meenan, R.T., Bachman, D.J., O'Keeffe Rosetti, M.C.: Risk adjustment using automated ambulatory pharmacy data: the RxRisk model. Med. Care **41**(1), 84–99 (2003)
13. Fleishman, J.A., Cohen, J.W., Manning, W.G., Kosinsk, M.: Using the SF-12 health status measure to improve predictions of medical expenditures. Med. Care **44**, 54–63 (2006)
14. Fox, J.: Applied Regression Analysis and Generalized Linear Models. Sage, Thousand Oaks (2008)
15. Garber, A.M., Macurdy, T.E., McClellan, M.B.: Persistence of medicare expenditures among elderly beneficiaries. In: Garber, A.M. (ed.) Frontiers in Health Policy, MIT Press, Cambridge, MA, pp. 153–180 (1998)
16. Gilmer, T., Kronick, R., Fishman, P., Ganiats, T.G.: The medicaid R x model: pharmacy-based risk adjustment for public programs. Med. Care **39**(11), 1188–1202 (2001)
17. Hughes, J.S., Averill, R.F., Eisenhandler, J., Goldfield, N.I., Muldoon, J., Neff, J.M., Gay, J.C.: Clinical Risk Groups (CRGs): a classification system for risk-adjusted capitation-based payment and health care management. Med. Care **42**(1), 81–90 (2004)
18. Jones, A.M.: Models for health care. Working paper, University of York. http://www.york.ac.uk/res/herc/documents/wp/10_01.pdf (2010). Accessed 13 June 2013
19. Kronick, R., Gilmer, T., Dreyfus, T., Lee, L.: Improving health-based payment for medicaid beneficiaries: CDPS. Health Care Financ. Rev. **21**(3), 29–64 (2000)
20. Lamers, L.M., van Vliet, R.C.J.A.: Multiyear diagnostic information from prior hospitalizations as a risk-adjuster for capitation payments. Med. Care **34**, 549–561 (1996)
21. Lamers, L. M.: Capitation payments to competing Dutch sickness funds based on diagnostic information from prior hospitalizations. Ph.D. Dissertation, Erasmus University Rotterdam, Rotterdam (1997)
22. Lamers, L.M.: Health-based risk adjustment: is inpatient and outpatient diagnostic information sufficient? Inquiry **38**(4), 423–431 (2001)
23. Lamers, L.M., van Vliet, R.C.J.A.: Health-based risk adjustment improving the pharmacy-based cost group model to reduce gaming possibilities. Euro. J. Health Econ. **4**, 107–114 (2003)
24. Lamers, L.M., van Vliet, R.C.J.A.: The pharmacy-based cost group model: validating and adjusting the classification of medications for chronic conditions to the Dutch situation. Health Policy **68**, 113–121 (2004)
25. Manning, W.G., Mullahy, J.: Estimating log models: to transform or not to transform? J. Health Econ. **20**, 461–494 (2001)
26. Manning, W.G., Busa, A., Mullahy, J.: Generalized modelling approaches to risk adjustment of skewed outcomes data. J. Health Econ. **24**, 465–488 (2005)
27. Monheit, A.C.: Persistence in health expenditures in the short run: prevalence and consequences. Med. Care **41**(7), 53–64 (2003)
28. McIntyre, S.H., Montgomery, D.B., Srinivasan, V., Weitz, B.A.: Evaluating the statistical significance of models developed by stepwise regression. J. Mark. Res. **20**, 1–11 (1983)
29. Mihaylova, B., Briggs, A., O'Hagan, A., Thompson, S.G.: Review of statistical methods for analysing healthcare resources and costs. Health Econ. **20**, 879–916 (2011)
30. Newhouse, J.P.: Reimbursing health plans and health providers: efficiency in production versus selection. J. Econ. Lit. **XXXIV**, 1236–1263 (1996)
31. Pindyck, R.S., Rubinfeld, D.L.: Econometric models and economic forecasts. McGraw-Hill, New York City (1998)
32. Pope, G.C., Ellis, R.P., Ash, A.S., Liu, C.F., Ayanian, J.Z., Bates, D.W., Burstin, H., Iezzoni, L.I., Ingber, M.J.: Principal inpatient diagnostic cost group model for medicare risk adjustment. Health Care Financ. Rev. **21**(3), 93–118 (2000)
33. Pope, G.C., Kautter, J., Ellis, R.P., Ash, A.S., Ayanian, J.Z., Iezzoni, L.I., Ingber, M.J., Levy, J.M., Robs, J.: Risk adjustment of medicare capitation payments using the CMS-HCC model. Health Care Financ. Rev. **25**(4), 119–141 (2004)
34. Powers, C.A., Meyer, C.M., Roebuck, M.C., Vaziri, B.: Predictive modeling of total healthcare costs using pharmacy claims data. A comparison of alternative econometric cost modeling techniques. Med. Care **43**, 1065–1072 (2005)
35. Prinsze, F.J., van Vliet, R.C.J.A.: Health-based risk adjustment: improving the pharmacy-based cost group model by adding diagnostic cost groups. Inquiry **44**(4), 469–480 (2007)
36. Schokkaert, E., van de Voorde, C.: Risk selection and the specification of the conventional risk adjustment formula. Eur. J. Health Econ. **23**, 1237–1259 (2004)
37. Schokkaert, E., van de Voorde, C.: Incentives for risk selection and omitted variables in the risk adjustment formula. Ann. Econ. Stat. **83**(84), 327–351 (2006)
38. Schokkaert, E., van de Voorde, C.: Direct versus indirect standardization in risk adjustment. J. Health Econ. **28**, 361–374 (2009)
39. Schut, F.T., van de Ven, W.P.M.M.: Uitvoering AWBZ door zorgverzekeraars onverstandig. ESB **95**(4591), 486–489 (2010)
40. Stam, P.J.A.: Testing the effectiveness of risk equalization models in health insurance. Ph.D. Dissertation, Erasmus University Rotterdam, Rotterdam (2007)

41. Stam, P.J.A., van de Ven, W.P.M.M.: Risicoverevening in de zorgverzekering: Een evaluatie en oplossingsrichtingen voor verbetering. Research Report, iBMG, Erasmus University Rotterdam, Rotterdam (2006)
42. Stam, P.J.A., van de Ven, W.P.M.M.: De harde kern in risicoverevening. ESB. February, 104-7 (2008)
43. Stam, P.J.A., van Vliet, R.C.J.A., van de Ven, W.P.M.M.: Diagnostic, pharmacy-based, and self-reported health measures in risk equalization models. Med. Care **48**, 448–457 (2010)
44. Thompson, M.L.: Selection of variables in multiple regression: part 1. a review and evaluation. Int. Stat. Rev. **46**, 1–19 (1978)
45. Veazie, P.J., Manning, W.G., Kane, R.L.: Improving risk adjustment for medicare capitated reimbursement using nonlinear models. Med. Care **41**, 741–752 (2003)
46. van Kleef, R.C., van Vliet, R.C.J.A.: Prior use of durable medical equipment as a risk adjuster for health-based capitation. Inquiry **47**, 1–16 (2010)
47. van Kleef, R.C., van Vliet, R.C.J.A.: Improving risk equalization using multi-year high cost as a health indicator. Med. Care **50**, 140–144 (2012)
48. van Kleef, R.C., van Vliet, R.C.J.A., van de Ven, W.P.M.M.: Risicoverevening tussen zorgverzekeraars: Kwantificering modelverbeteringen 1993–2011. TSG **90**, 312–326 (2012)
49. van Kleef, R.C, van Vliet, R.C.J.A., van de Ven, W.P.M.M.: Risicoverevening 2012. Een analyse van voorspelbare winsten en verliezen op subgroep niveau. Research report, iBMG, Erasmus University Rotterdam, Rotterdam (2012)
50. van Kleef, R.C., van Vliet, R.C.J.A., van de Ven, W.P.M.M.: Diagnosis-based cost groups in risk adjustment: The effects of including outpatient diagnoses. Research report, iBMG, Erasmus University Rotterdam, Rotterdam (2012)
51. van de Ven, W.P.M.M., Ellis, R.P.: Risk adjustment in competitive health plan markets. In: Cutler, A., Newhouse, J.P. (eds) Handbook of health economics, pp. 755-845. Elsevier Science B.V., Amsterdam (2000)
52. van de Ven, W.P.M.M., Schut, F.T.: Guaranteed access to affordable coverage in individual health insurance markets. In: Glied, S., Smith, P. (eds.) The Oxford Handbook of Health Economics, pp. 380–404. Oxford University Press, Oxford (2011)
53. van Vliet, R.C.J.A., van de Ven, W.P.M.M.: Towards a capitation formula for competing health insurers: an empirical analysis. Soc. Sci. Med. **34**, 1035–1048 (1992)
54. van Vliet, R.C.J.A., van de Ven, W.P.M.M.: Capitation payments based on prior hospitalizations. Health Econ. **2**, 177–188 (1993)
55. Ware Jr, J.E., Kosinski, M., Keller, S.D.: A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. Med. Care **34**, 220–233 (1996)