Evaluating and Improving the Predictive Performance of Risk Equalization Models in Health Insurance Markets

Het evalueren en verbeteren van de voorspelkracht van risicovereveningsmodellen op zorgverzekeringsmarkten

Suzanne H.C.M. van Veen

Evaluating and Improving the Predictive Performance of Risk Equalization Models in Health Insurance Markets

Het evalueren en verbeteren van de voorspelkracht van risicovereveningsmodellen op zorgverzekeringsmarkten

S.H.C.M. (Suzanne) van Veen

© S.H.C.M. van Veen, 2016. All rights reserved.
No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by means, electronically, mechanically, by photocopying, recording, or otherwise, without the prior written permission of the author.

Chapters based on articles published in peer-reviewed scientific journals and reprinted with kind permission of SAGE publications, Springerlink, and John Wiley & Sons Ltd.

Cover design: Optima Grafische Communicatie (ww.ogc.nl)
Layout and printed by: Optima Grafische Communicatie (www.ogc.nl)

ISBN: 978-94-6169-781-3

Evaluating and Improving the Predictive Performance of Risk Equalization Models in Health Insurance Markets

Het evalueren en verbeteren van de voorspelkracht van risicovereveningsmodellen op zorgverzekeringsmarkten

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

7 april 2016 om 13.30 uur

door

Suzanne Hendrika Catharina Maria van Veen
geboren te Delft

Erasmus University Rotterdam

PROMOTIECOMMISSIE

Promotor:

prof.dr W.P.M.M. van de Ven

Overige leden:

prof.dr. E.K.A. van Doorslaer
prof.dr. E. Schokkaert
prof.dr. J. Boone

Copromotoren:

dr. R.C. van Kleef
dr. R.C.J.A. van Vliet

Contents

Chapter 1 Introduction **15**

§ 1.1 Background 17
§ 1.2 Relevance of an adequate risk equalization model 19
§ 1.3 Evaluation criteria for risk equalization models 21
§ 1.4 Research questions and relevance 23
§ 1.5 Data and methods 27
§ 1.6 Goal and structure of this thesis 30

Part I Evaluating the Predictive Performance of Risk Equalization Models

Chapter 2 A Taxonomy and Review of Measures-of-Fit **35**

§ 2.1 Introduction 38
§ 2.2 New contribution 39
§ 2.3 Theoretical framework 40
§ 2.4 Method 41
§ 2.5 Results 43
§ 2.6 Conclusions and discussion 54

Appendix 2.1: Search strategy 59
Appendix 2.2: References of the studies included in our review 61
Appendix 2.3: Description of the measures-of-fit 66

Chapter 3 Estimating the Potential Selection Profits **85**

§ 3.1 Introduction 88
§ 3.2 How to estimate the potential selection profits? 91
§ 3.3 Empirical analysis 97
§ 3.4 Empirical findings 103
§ 3.5 Conclusions 108
§ 3.6 Discussion 109

Appendix 3.1: Descriptive statistics 115
Appendix 3.2: Models' predictive performance on the full sample 118
Appendix 3.3: Relationship between the R^2 and CPM 119

Part II Improving the Predictive Performance of Risk Equalization Models

Chapter 4 Cost and Diagnostic Information from Multiple Prior Years 123

§ 4.1 Introduction 126
§ 4.2 Data and methods 129
§ 4.3 Results 137
§ 4.4 Conclusions 142
§ 4.5 Discussion 142

Appendix 4.1: Definition of the risk adjusters 149
Appendix 4.2: The Mean Prediction Error for some evaluation-groups 150

Chapter 5 Interaction Terms between Existing Risk Adjusters 153

§ 5.1 Introduction 156
§ 5.2 Data and methods 158
§ 5.3 Results 165
§ 5.4 Conclusions and discussion 168

Appendix 5.1: Detailed descriptive statistics (I) 173
Appendix 5.2: Detailed descriptive statistics (II) 175
Appendix 5.3: Regression tree parameters 177
Appendix 5.4: Results of the sensitivity analysis 181
Appendix 5.5: The Mean Prediction Error for some evaluation-groups 182

Chapter 6 Residual Expenses from Multiple Prior Years 185

§ 6.1 Introduction 188
§ 6.2 Data and methods 190
§ 6.3 Results 196
§ 6.4 Conclusions and discussion 209

Appendix 6.1: Persistence in residual expenses over a four-year time period 217
Appendix 6.2: Detailed risk characteristics of the persistently under-compensated groups 219
Appendix 6.3: Results of the sensitivity analysis 220
Appendix 6.4: The predictive performance for some evaluation-groups 221

Chapter 7 Conclusions and Discussion **225**

§ 7.1 How to evaluate the predictive performance of risk equalization models? 227

§ 7.2 To what extent can the predictive performance of a morbidity-based risk equalization model be improved by three new methods? 230

§ 7.3 Discussion on evaluating risk equalization models 233

§ 7.4 Discussion on improving risk equalization models by the three new methods 241

§ 7.5 Recommendations 244

§ 7.6 Directions for further research 247

General Appendices **251**

General Appendix 1: Risk equalization in the Netherlands 253

General Appendix 2: Definition of the evaluation-groups 257

Abbreviations **261**

References **265**

Samenvatting **277**

Dankwoord **289**

Curriculum Vitae **295**

About the author **305**

LIST OF PUBLICATIONS AND SUBMISSIONS

Chapters 2 through 6 are based on the following articles:

Chapter 2
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 Risk equalization Models. *Medical Care Research and Review, 72*(2), 220-243

Chapter 3
van Veen, S.H.C.M., van Kleef, R.C., van de Ven, W.P.M.M., van Vliet, R.C.J.A. (2015). Evaluating Risk Equalization Models by estimating the Potential Selection Profits. *Submitted for publication.*

Chapter 4
van Veen, S.H.C.M., van Kleef, R.C., van de Ven, W.P.M.M., van Vliet, R.C.J.A. (2015). Improving the Prediction Model used in Risk Equalization: Cost and Diagnostic Information from Multiple Prior Years. *European Journal of Health Economics, 16*(2), 201-218.

Chapter 5
van Veen, S.H.C.M., van Kleef, R.C., van de Ven, W.P.M.M., van Vliet, R.C.J.A. (2015). Exploring the Predictive Power of Interaction Terms in a Sophisticated Risk Equalization Model using Regression Trees. *Submitted for publication.*

Chapter 6
van Veen, S.H.C.M., van Kleef, R.C., van de Ven, W.P.M.M., van Vliet, R.C.J.A. (2015). Exploring Persistent Under-compensations under a Morbidity-based Risk Equalization Model: Evidence from the Netherlands. *Submitted for publication.*

Chapter 1

Introduction

§ 1.1 BACKGROUND

Over the past decades, several countries around the world – e.g. Belgium, Germany, Israel, the Netherlands, Switzerland, and the U.S. – have introduced (principles of) regulated competition[1] among health insurers (van de Ven et al., 2007; Kautter et al., 2014). The ultimate goal of such regulated competition is to stimulate efficiency and responsiveness to consumers' preferences, while at the same time universal financial access to health insurance is guaranteed (van de Ven & Schut, 2011). In order to achieve this goal, several crucial preconditions needs to be fulfilled (van de Ven & Schut, 2011; van de Ven et al., 2013). One of them is the implementation of an *adequate risk equalization (RE) scheme*.

RE is a regulatory scheme of (prospective) risk-adjusted payments that are provided to insurers. The regulator (e.g. government) organizes an 'RE Fund' with mandatory contributions from taxes, employers, insurers, or consumers. The money in this fund is allocated among insurers by means of risk-adjusted payments for each enrollee. In this way, insurers are compensated for predictable variation in individuals' healthcare expenses. The risk-adjusted payments are based on an *RE-model*, which is a prediction model that uses various risk factors to predict individuals' healthcare expenses. For example, an insurer with an above-average proportion of elderly in his portfolio receives higher compensations than an insurer with a below-average proportion of elderly, because elderly are expected to have above-average expenses.

All countries with an RE scheme also have introduced premium regulation[2] (van de Ven et al., 2013; Kautter et al., 2014). In the presence of premium regulation, the goal of RE is to mitigate financial incentives for risk selection and thereby to achieve a level playing field[3]. Risk selection can be defined as "actions (other than risk rating per product) by consumers and insurers with the intention and/or the effect that solidarity [i.e. the intended pooling of low-cost and high-cost individuals] is not fully achieved" (van Kleef et al., 2013a; Newhouse, 1996). Risk selection is a potential threat to solidarity, efficiency, and quality of care, as discussed in more detail below. The extent to which an RE-model mitigates financial incentives for risk selection depends on the *predictive performance* of the model. 'Predictive performance' refers to the statistical fit of the model; i.e. how well the model predicts

[1] The elements that have been introduced in these countries are based on Enthoven's Managed Competition Model (Enthoven, 1988).

[2] Premium regulation is used often in the extreme form of community-rating, implying that an insurer must charge the same premium for the same product, e.g. this is the case in Belgium, Germany, the Netherlands, and Switzerland, or premium regulation is used in the form of a premium bandwidth, e.g. in the U.S. premium variation by age is constrained to 3:1 and by smoking to 1.5:1 (Kautter et al., 2014; van de Ven et al., 2013).

[3] A level playing field is obtained when all insurers are compensated for the expected expenses of the risk composition of their portfolio, given the average efficiency across insurers in providing services.

expenses for individuals and/or groups. An RE-model that adequately predicts expenses for any selective group of interest completely removes financial incentives for risk selection.

Over the past two decades, there has been a widespread interest in *improving* the predictive performance of RE-models. Among researchers and policymakers it has been well understood that simple demographic RE-models do not predict expenses adequately for many specific groups of interest. Hence, additional risk adjusters that are better indicators for individuals' health status than solely age and gender were necessary to reduce financial incentives for risk selection. This resulted in the development of morbidity-based risk adjusters based on inpatient or outpatient diagnostic information, pharmacy information, or prior years' cost information (e.g. Adams et al., 2002; Buchner et al., 2013; Ellis & Ash, 1995; Fishman et al., 2003; Hughes et al., 2004; van Kleef & van Vliet, 2012; Kronick et al., 2000; Pope et al., 2000a). Consequently, the predictive performance of RE-models that are used in several countries has been improved considerably (Buchner et al., 2013; Kautter et al., 2014; van Kleef et al., 2014; van de Ven et al., 2007). However, an important question in all countries with RE is still how to further improve the predictive performance because research has shown that even sophisticated morbidity-based RE-models under- and over-predict expenses for specific groups in the population, despite the model improvements over the past decades (e.g. Behrend et al., 2007; van Kleef et al., 2014; Payne et al., 2000).

In contrast to the vast amount of literature paying attention to improving the predictive performance of RE-models, *evaluating* model's performance has been understudied. The literature provides little guidance on how to evaluate RE-models, while numerous empirical evaluations have been conducted. The evaluation of model's performance is important because it is used to monitor financial incentives for risk selection and it is a crucial aspect of each study examining a potential model improvement. The results of empirical evaluations serve as a basis for deciding on the design of the RE-model, which requires a good understanding of the methods that are used for the evaluation of model's performance.

Against this background, this thesis focusses on the evaluation of the predictive performance of RE-models and addresses the question to what extent the predictive performance of one of the most sophisticated RE-models that are used around the world could be improved by three new methods. The three methods examined in this thesis are: additional risk adjusters based on cost and/or diagnostic information from multiple prior years, additional interaction terms between existing risk adjusters, and additional risk adjusters or interaction terms based on prior years' residual expenses. The central question of this thesis reads:

> *How to evaluate the predictive performance of risk equalization models and to what extent can the predictive performance of a morbidity-based risk equalization model be improved by three new methods?*

Before formulating the specific research questions of this thesis, the next section first describes in more detail why RE is a crucial precondition in health insurance markets. Thereafter, the relevance of an adequate RE-model is addressed by describing potential threats of risk selection that are present when an imperfect RE-model is used. Section 1.3 describes different evaluation criteria for RE-models and motivates why this thesis focusses on the predictive performance. Furthermore, this section explains what is exactly meant with 'predictive performance' and when an RE-model is considered to be adequate.

§ 1.2 RELEVANCE OF AN ADEQUATE RISK EQUALIZATION MODEL

§ 1.2.1 Why is risk equalization crucial in health insurance markets?

In *unregulated* competitive health insurance markets, insurers adjust the premium to individuals' expected expenses[4] per contract in order to be financially viable (i.e. equivalence principle). Given the large variation in expected expenses among individuals, risk-rated premiums may make health insurance unaffordable for those who need healthcare most[5]. Furthermore, insurers may have financial incentives to engage in risk selection, because for some individuals it may be impossible or too costly to calculate risk-rated premiums. This may lead to refusal of certain individuals from enrolling or exclusion of pre-existing medical conditions from coverage. Thus, to achieve society's goal of reconciling efficiency and universal financial access to health insurance it is needed to enforce some regulations in the market.

A straightforward way to make health insurance affordable for high-risk individuals in a (competitive) health insurance market is to organize cross-subsidies from low-risk to high-risk individuals (Enthoven, 1988; van de Ven & Ellis, 2000). An open enrolment requirement[6] guarantees access to health insurance but is itself non-restrictive on risk rating. There are several types of cross-subsidies that can be used: explicit subsidies in the form of premium-based subsidies, risk-adjusted subsidies, or ex-post compensations; and implicit subsidies in the form of a guaranteed renewability requirement[7] or premium-rate restrictions (van de Ven & Schut, 2011).

---

[4] In statistical terms, 'expected expenses' refer to '*predicted* expenses', meaning those expenses that can be estimated by a prediction model. Throughout this thesis, 'expected expenses' and 'predicted expenses' are used interchangeable.

[5] If an insurer fully adjusts the premium to individuals' expected expenses, the premium for a high-cost individual may be the *500-fold* of the premium for a low-cost individual per year (Eijkenaar et al., 2015).

[6] An open enrollment obligates insurers to accept any individual for a standard benefit package.

[7] A guaranteed renewability requirement obligates the insurer to renew the contract at the end of each contract period for a standard premium and with a standard benefit package; however, this requirement does not impose restrictions on insurers to contract new enrollees.

Van de Ven & Schut (2011) have concluded that the first and best form of regulation in a (competitive) health insurance market to reconcile efficiency and universal financial access to health insurance is the implementation of an RE scheme; i.e. a regulatory scheme of explicit (prospective) risk-adjusted payments to insurers. Premium-based subsidies and ex-post compensations may not provide optimal incentives for efficiency and therefore, may lead to an inevitable trade-off between efficiency and affordability. Premium-rate restrictions may guarantee affordability but provides financial incentives for risk selection, because differences in individuals' expected expenses cannot be reflected in the premium, which inevitably leads to a trade-off between affordability and risk selection. Guaranteed renewability is also not preferred, because for high-risk individuals it restricts a free periodic choice of insurer, which is another crucial precondition that needs to be fulfilled in a regulated (competitive) health insurance market (van de Ven et al., 2013).

All countries with an RE scheme have implemented also other regulations, such as an open enrolment requirement and premium-rate restrictions. The open enrolment requirement and premium regulation guarantee universal financial access to health insurance with a standard coverage. *Without* RE, these two regulations would provide large financial incentives to engage in risk selection because then insurers are not compensated for predictable variation in individuals' healthcare expenses. In this context, a particular goal of RE is to mitigate these financial incentives for risk selection. The better the RE-model compensates insurers for predictable variation in individuals' healthcare expenses, the less severe the trade-off between efficiency and risk selection. Hence, an *adequate RE-model* is the only strategy that can reconcile efficiency and universal financial access to health insurance, without creating financial incentives for risk selection (van de Ven & Schut, 2011).

§ 1.2.2 Potentially threating effects of risk selection

One of the most important concerns of an imperfect RE-model is the financial incentives for risk selection that are provided. There are several possible forms of risk selection that in most countries generally are not prohibited by law. Examples are service-level distortion (e.g. not contracting the first-best healthcare provider for a specific treatment), providing poor service such as a long query-response time, selective marketing, or group-contracts (Baumgartner & Busato, 2012; Beck et al., 2003; Newhouse, 1996; van Kleef et al., 2013a; von Wyl & Beck, 2015; van de Ven & Ellis, 2000).

Eventually, these actions may undermine solidarity, reduce efficiency, and distort quality of care (Baumgartner & Busato, 2012; Beck et al., 2003; Frank et al., 1998; Newhouse, 1996; von Wyl & Beck, 2015; van de Ven & Ellis 2000). First, solidarity is undermined when risk selection leads to segmentation of high-risk individuals and low-risk individuals. High-risk individuals may end up in different health insurance products and/or by different insurers than low-risk individuals, which may lead to a high premium for these high risks and eventually, an insurer with a relatively large proportion high-risk individuals

may go bankrupt ('death spiral'). Second, risk selection may reduce efficiency when the expected benefits of risk selection exceed the expected benefits on investing in efficiency improvements. In addition, investing in risk selection – and not in efficiency or quality of care – can be considered a welfare loss to society. Third, insurers (and/or providers) may have disincentives to respond to the patients' preferences, to invest in quality of care, and to acquire the best reputation for treating specific patient groups.

Recently, some studies have found evidence of risk selection in practice (Newhouse et al., 2012, 2015; Shmueli & Nissan-Engelcin, 2013; van Kleef et al., 2013a; von Wyl & Beck, 2015). This emphasizes the importance of further improving the predictive performance of the RE-models that are used in order to reduce the financial incentives for risk selection and so, reduce its potentially threating effects on solidarity, efficiency, and quality of care.

§ 1.3 EVALUATION CRITERIA FOR RISK EQUALIZATION MODELS

§ 1.3.1 Different evaluation criteria

To decide on the design of the RE-model to be used in practice requires incorporating different evaluation criteria. Van de Ven & Ellis (2000) grouped numerous criteria into three broad categories: appropriateness of incentives, fairness, and feasibility, which are summarized in Table 1.1. This list is *not* exhaustive and the criteria are *not* necessarily presented in order of relative importance.

The different evaluation criteria may be related. For example, a risk adjuster with a high predictive power may lead to higher compensations for high-risk individuals than for low-risk individuals (i.e. fairness) and so, may mitigate financial incentives for risk selection (i.e. provide appropriate incentives). Moreover, the different evaluation criteria may sometimes be in conflict. As a result, incorporating all evaluation criteria may make decision-making about the design of the RE-model complex because of (inevitable) trade-offs. For example, including a risk adjuster based on prior years' expenses in an RE-model may increase the predictive performance and so, mitigates financial incentives for risk selection; however, this risk adjuster may reduce efficiency. Furthermore, feasibility imposes constraints on the risk adjusters that can be included in RE-models, e.g. information that is available for all individuals in the population may be 'only' a proxy of individuals' health status, such as prior years' expenses, while complete medical information about individuals' health status may not be routinely available or it will require a large amount of money and/or time to collect and analyze for the purpose of RE. In practice, policymakers have to decide how to value the different evaluation criteria, which may result in different outcomes in different countries.

Table 1.1: A list of important evaluation criteria for risk adjusters in risk equalization models

	Definition
Appropriateness of incentives	A risk adjuster should provide appropriate incentives, which means correcting financial incentives for risk selection and not stimulating inefficient behavior or low quality of care[a]. For instance, this criterion implies that a risk adjuster should have *predictive power* in order to mitigate financial incentives for risk selection, and a risk adjuster should not be subject to manipulation or fraud in order to stimulate efficient behavior[b].
Fairness	This criterion is concerned with value judgements about the risk factors for which the regulator desires compensation (i.e. the S-type risk factors)[d] and the risk factors for which it does not (i.e. the N-type risk factors)[e], and whether the risk adjuster adjusts expenses for those (groups of) individuals for whom it is desired to compensate insurers. Further, this criterion is concerned with the value judgment that the compensation should be higher for individuals who are sicker than those who are less sick[a].
Feasibility	It should be administratively feasible to obtain information from all individuals in the relevant population without undue time and money. Further, the time lag between data collection and when it is feasible to use for payments should not be too long[b,c]. Furthermore, this criterion is concerned with the acceptability, validity, and reliability of a risk adjuster. *Acceptability* means that the risk adjuster should satisfy privacy protection laws and clinical credibility. Clinical credibility means that the risk adjuster is based on an accepted classification algorithm that leads to a higher payment for those (groups of) individuals with a higher need for healthcare services [c]. *Validity* means that a risk adjuster should predict the need for healthcare services and define relatively homogeneous groups. This implies that the risk adjuster should identify chronic or other conditions that are predictable, rather than acute health episodes that represent insurers' financial risk[b,c]. *Reliability* implies that a risk adjuster should be measured without measurement error[b].

Footnotes Table 1.1:

a. See van de Ven & Ellis, 2000.
b. See van Vliet & van de Ven, 1993.
c. See Kautter et al., 2014.
d. Examples of S(subsidy)-type risk factors: age, gender, health status.
e. Examples of N(on subsidy)-type risk factors: practice variation, inefficiency in provision of services.

§ 1.3.2 Predictive performance

A central theme in this thesis is the *predictive performance* of RE-models, which is only one of all evaluation criteria for RE-models. The motivation to focus on this criterion is that it is one of the most important ones, because this criterion can be used to quantify the extent to which an RE-model achieves its policy goal (i.e. reducing financial incentives for risk selection). A higher predictive performance is generally desirable, because this may lead to reduced financial incentives for risk selection.

The evaluation of the predictive performance of RE-models consists of applying one or more measures-of-fit to quantify residual expenses for individuals and/or groups; the evaluation of RE-models does *not* focus on hypothesis testing or causal interpretation of the risk adjusters. 'Residual expenses' refer to the difference between observed expenses and predicted expenses. In the literature, also the term 'prediction error' has been used. Furthermore, positive residual expenses mean that predicted expenses fall below observed expenses, which can be interpreted as an 'under-compensation' or 'under-prediction'. Negative residual expenses mean that predicted expenses exceed observed expenses, which can be interpreted as an 'over-compensation' or 'over-prediction'. Throughout this thesis,

'residual expenses', 'prediction error', 'under- and over-compensation', and 'under- and over-prediction' are used interchangeable. Furthermore, throughout this thesis observed expenses are used as the reference point for calculating residual expenses. Ideally, normative expenses are used as the reference point rather than observed expenses (Stam, 2007). Normative expenses are those expenses related to risk factors for which the regulator wants compensation (i.e. the S-type risk factors, such as individuals' health status) and *not* those related to risk factors for which the regulator does not want compensation (i.e. the N-type risk factors, such as inefficiencies in provision of services). However, normative expenses are not used in practice because of several practical limitations; for instance, the availability of information to estimate normative expenses and the complexity of distinguishing S-type from N-type cost variation (Schokkaert & van de Voorde, 2004, 2006, 2009). Consequently, an unknown part of residual expenses may be due to N-type cost variation.

An RE-model is considered to be adequate when average residual expenses for any selective group that is of interest to the regulator are (close to) zero. If an RE-model adequately predicts expenses for selective groups of interest, the differences in the (out-of-pocket) premiums across insurers would reflect the net differences in efficiency and administrative costs of insurers for providing services and contracting or delivering healthcare services; and *not* differences in the risk composition of insurers' portfolio. In this situation, there are incentives for efficiency and responsiveness to consumers' preferences, while universal access to health insurance is guaranteed.

It is worth noting that an RE-model does not have to, and cannot, predict expenses adequately for each individual in the population, because observed expenses include a random component (i.e. unpredictability). Improving the predictive performance of RE-models focusses on obtaining adequate predictions of expenses for selective groups of interest.

§ 1.4 RESEARCH QUESTIONS AND RELEVANCE

To answer the central question of this thesis: *"How to evaluate the predictive performance of risk equalization models and to what extent can the predictive performance of a morbidity-based risk equalization model be improved by three new methods?"*, this thesis is divided into two parts with the first part focusing on evaluating the predictive performance of RE-models and the second part on improving the predictive performance of RE-models. Each part consists of research questions that are examined in separate chapters.

PART I – EVALUATING THE PREDICTIVE PERFORMANCE OF RISK EQUALIZATION MODELS

As noted earlier, to know the extent to which a given RE-model mitigates financial incentives for risk selection, it is required to evaluate model's predictive performance. Furthermore, evaluating model's performance is an essential element of each study investigating newly-developed risk adjusters. Such empirical evaluations generally consist of a comparison of the performance of a benchmark model, e.g. the RE-model that is used in practice, to one or more alternative RE-models, with the aim to determine the extent to which the statistical fit of the model can be improved and so, which RE-model should be used in practice.

Given the amount of empirical evaluations that has been performed over the past decades, one would expect there to be a substantial amount of literature paying attention to the way RE-models should be evaluated, which may lead to consistency across studies in the evaluation method. However, numerous measures-of-fit have been applied in many different ways, without systematic attention on evaluating the performance of RE-models. Research has shown that different measures-of-fit and different analytic methods may lead to different conclusions about model's performance and hence may lead to different decisions concerning the RE-model to be used (e.g. Fishman et al., 2003; Powers et al., 2005; Stam et al., 2010b). The first part of this thesis contributes to the literature by providing a taxonomy of measures-of-fit for RE-models and discussing important variations in the method of applying these measures. The aim is to formulate general principles on how to evaluate model's predictive performance. Therefore, the first research question reads:

> *Question 1: Which measures-of-fit have been used for evaluating risk equalization models and how have these measures been applied?*

Chapter 2, which addresses this question, argues that the only appropriate method to measure the extent to which an RE-model mitigates financial incentives for risk selection is to evaluate the statistical fit for *selective groups*. Several studies have applied this evaluation method by examining average residual expenses per RE-model for each group of interest (e.g. Ash & Byrne-Logan, 1998; Ash et al., 2005; van Kleef et al., 2012a, 2012b, 2013b; Pope et al., 2000a; van Veen et al., 2015b). However, with this conventional evaluation method it is possible that conflicting results are obtained because RE-models may improve the prediction of expenses for some groups, while deteriorating the prediction for others. It is a challenge to develop an evaluation method that estimates the potential selection profits for multiple groups simultaneously. Such an evaluation method may be very helpful to monitor to what extent an RE-model achieves its policy goal and to decide which RE-model may lead to the largest *overall* reduction in the financial incentives for risk selection for different

groups and so, should be used in practice. The lack of such an evaluation method served as the motivation for the second research question of this thesis, which reads:

> *Question 2: How to estimate the potential selection profits for multiple groups simultaneously under a risk equalization model?*

Regarding question 2, three different methods are developed and empirically tested in order to determine whether it matters which method is used for estimating the potential selection profits under an RE-model. The crux for estimating these profits is to create *mutually exclusive groups*. Overlap between groups may especially occur for groups based on the presence of chronic conditions or prior healthcare utilization. Since mutually exclusive groups can be defined in different ways, three alternative methods were examined. In addition, these methods are compared with a relatively simple method of aggregating average residual expenses for multiple overlapping groups, as used in previous studies (van Kleef et al., 2012a; Ash et al., 2005). Aggregating average residual expenses for overlapping groups yields a biased estimate of the potential selection profits in absolute monetary terms, because of double-counting of individuals who occur in multiple groups. The purpose of comparing this method to the developed methods is to investigate to what extent this relatively simple method biases the estimates of the potential selection profits and hence leads to another conclusion about which RE-model should be used. All methods used the same set of predefined overlapping selective groups of interest as the starting point.

PART II – IMPROVING THE PREDICTIVE PERFORMANCE OF RISK EQUALIZATION MODELS

Sophisticated morbidity-based RE-models that are used in several countries – e.g. Belgium, Germany, the Netherlands, and the U.S. – use an advanced set of risk adjusters based on demographic information, diagnostic information, and/or cost information from previous years (Buchner et al., 2013; Kautter et al., 2014; van Kleef et al., 2014; van de Ven et al., 2007). This thesis examines three potentially relevant methods to improve the predictive performance of such sophisticated morbidity-based RE-models. All three methods are based on the principle of extending the set of risk adjusters by developing new types of risk adjusters or interaction terms between existing ones. Furthermore, they are based on information already available in the administrative files of (Dutch) insurers. An advantage of exploring potentially relevant risk adjusters or interaction terms that are based on available information is that there are no additional costs for data collection, which may make it possible to implement them with relatively low administrative costs.

A first potentially relevant model improvement examined here is the usage of cost and/or diagnostic information from multiple prior years. The motivation to explore the predictive power of this type of information is that existing RE-models do not fully exploit the information that is available in insurers' administrative files from multiple prior years: e.g. some RE-models 'only' use risk adjusters based on information from one prior year but *not* multiple years, as is the case in Belgium, Germany, Israel, Switzerland, and the U.S. (Buchner et al, 2013; Kautter et al, 2014; van de Ven et al., 2007), or a risk adjuster based on cost information from multiple prior years but *not* diagnostic information from multiple prior years – i.e. pharmacy-based cost groups (PCGs) or diagnostic cost groups (DCGs) – as is the case in the Netherlands (van Kleef & van Vliet, 2012). An innovative approach is applied to explore the predictive power of a large array of multiyear cost-based and diagnostic-based risk adjusters. The third research question is therefore:

> *Question 3: To what extent can the predictive performance of a morbidity-based risk equalization model be improved by including risk adjusters based on cost and/or diagnostic information from multiple prior years?*

Besides multi-year cost and/or diagnostic information, a second potentially relevant method to improve an RE-model is the inclusion of interaction terms between existing risk adjusters. Interaction terms have been applied moderately in RE-models, while these may be especially useful for predicting expenses of some selective groups of interest, such as individuals with co-morbidity (Pope et al, 2004). Given the large dataset analyzed here (N = ~16 million) and the complexity of the associations in the data, there could be theoretically more than one million interaction terms possible for sophisticated morbidity-based RE-models. To identify only those interactions that statistically significantly contribute to model's predictive performance, regression tree modelling is used. Regression tree modelling has been widely adopted in various scientific fields, but has been applied rarely within the field of RE (Robinson, 2008; Buchner et al., 2014). This study extends on previous research by several methodological improvements, aiming to explore to what extent models' performance can be improved when all statistically relevant interaction terms are included. Thus, the fourth question reads:

> *Question 4: To what extent can the predictive performance of a morbidity-based risk equalization model be improved by including interaction terms between existing risk adjusters?*

Besides the two previous methods, a third potentially relevant model improvement is to use prior years' residual expenses to define a risk adjuster or interactions terms for a group of individuals who are persistently under-compensated. For doing so, however, it is first necessary to identify individuals with persistent under-compensations in the population and explore the costs and risk characteristics of these individuals. In the literature, it is largely unknown whether there are individuals who are persistently under-compensated under a morbidity-based RE-model and what their costs and risk characteristics are. Therefore, a first main contribution to the literature is the identification of persistently under-compensated groups in the population over a three-year time period and providing insight into their costs and risk characteristics. A second contribution is to use this information to improve model's predictive performance, especially for those individuals with persistent under-compensations. Individuals who exhibit persistent under-compensations may be vulnerable to risk selection, because insurers systematically receive a lower risk-adjusted payment than the expected expenses of these individuals over several years. The fifth research question is therefore:

> *Question 5: To what extent can the predictive performance of a morbidity-based risk equalization model be improved by including a risk adjuster or interaction terms based on residual expenses from multiple prior years?*

Questions 3 through 5, addressed in Chapters 4 – 6, exclusively focus on the predictive power of the new risk adjusters and/or interaction terms. Chapter 7 reflects on other evaluation criteria for each of the three potentially relevant model improvements, aiming to provide a broader picture of the practical relevance of them.

§ 1.5 DATA AND METHODS

Research question 1 will be answered by conducting a systematic literature review. Research questions 2 through 5 are based on empirical analyses. These empirical analyses use Dutch administrative data and Dutch health survey data. These datasets are used to estimate alternative RE-models, with the Dutch RE-model being the benchmark. Consequently, the empirical findings are conditional on the data and this benchmark model. The next paragraphs will describe the datasets and general methods of the empirical analyses that are used to provide answers to the questions 2 through 5. The specific method that is used for each of these questions – i.e. the technical details and the innovative application of methods for the purpose of the study – is explained in detail in each separate chapter (see

Chapters 2 – 6). General appendix 1 (page 253) provides background information about the organization of the RE scheme in the Netherlands and explains in more detail the design of the Dutch RE-model.

§ 1.5.1 Administrative data

For the empirical analyses throughout this thesis, administrative data from the period 2006 to 2011 are used. These datasets have been used in practice for calculating the *actual* RE-models in the Netherlands.

The administrative dataset covers almost the entire Dutch population (N = ~16 million). For each individual, cost information, demographic information, and diagnostic information is available. Cost information includes annual observed expenses related to the services included in the basic benefit package in the respective year; e.g. hospital care, primary care, paramedical care, pharmaceuticals, durable medical equipment, transport in case of illness, obstetrical care, maternity care, and (long-term) mental healthcare. The basic benefit package may change from year to year, with new services included (e.g. new drugs) and/or some services excluded (e.g. some treatments by the dentist). Total observed expenses for all services included in the benefit package in the respective year, except (long-term) mental healthcare, are used for the empirical analyses. Mental healthcare expenses are excluded because in the Netherlands a separate RE-model with some other risk adjusters is used. Demographic and diagnostic information includes the risk adjusters in the Dutch RE-model.

§ 1.5.2 Health survey data

In addition to administrative data, health survey data from a (representative) sample of the Dutch population are used (N = ~8,000 for year 2008; N = ~16,000 for year 2010). This data are collected by a government agency, "Statistics Netherlands". The survey results are merged to the administrative data by using a unique identification key. Dutch privacy protection laws are followed for this procedure. The survey contains information on self-reported health status and healthcare utilization. To provide an answer to research questions 3 through 5, these data are used to evaluate the statistical fit for selective groups of interest, aiming to show the extent to which financial incentives for risk selection are mitigated (General appendix 2 defines the evaluation-groups used, see page 257). Survey data from 2008 are used for evaluating RE-models when administrative data from 2009 are used for estimating these models and survey data from 2010 are used for evaluating models when administrative data from 2011 are used for estimating these models. This is because when evaluating models' performance it is of interest to know how well the model predicts expenses for selective groups based on information that is known a priori the estimation-year. To provide an answer to research question 2, survey data are used to estimate the potential selection profits for multiple groups simultaneously under different RE-models.

§ 1.5.3 General methods for the empirical analyses

The Dutch RE-model has been developed over the past two decades. This RE-model includes the following risk adjusters: age interacted with gender (since the introduction of the model in 1993), region (since 1995), source of income interacted with age (since 1999), PCGs (since 2000), DCGs (since 2004), socioeconomic status interacted with age (since 2008), multiple-year high cost groups (MHC-groups, since 2012), and durable medical equipment groups (DME-groups, since 2014). Throughout this thesis, the Dutch RE-model of 2012, 2013, or 2014 is the benchmark, depending on the administrative data that were available at the time of performing the empirical analysis. As a result, the data and benchmark model are not kept constant.

The Dutch RE-model is estimated by Ordinary Least Squares (OLS), which is the conventional estimation technique for RE-models. The dependent variable is annual total observed expenses and the independent variables are the previously mentioned risk adjusters in the form of dummy variables. Ideally, normative expenses are the expenses to be compensated and not observed expenses (Stam, 2007). However, due to practical limitations normative expenses are not used in practice. General appendix 1 (see page 253) provides a more detailed description of the design of the Dutch RE-model. For the empirical analyses, the definitions of the risk adjusters that were available in the administrative data from each year, the total observed expenses in the respective year, and the estimation technique are taken as given. Discussing the pros and cons of the design the Dutch RE-model is beyond the scope of this thesis.

Though the empirical analyses use the Dutch RE-model as the benchmark, the findings of this thesis may also be relevant for countries with other RE-models. Chapter 7 will reflect on the findings by discussing what may be expected when each of the three methods examined here are applied while using an alternative benchmark. In principle, the methods used for investigating each of the three potential model improvements are generally applicable to any RE-model.

It is important to mention that the three potential model improvements are examined *separately*; and not sequentially. Via this way, they can be interpreted as three alternatives to improve the predictive performance.

§ 1.5.4 Implication of using Ordinary Least Squares for the evaluation methods

The assumption that an RE-model is estimated by OLS has an important implication on the method for evaluating the predictive performance. This is because in case of OLS average predicted expenses equals average observed expenses for all groups that are explicitly taken into account by the risk adjusters in the form of dummy variables. Consequently, any OLS-model, regardless the quality of this model, perfectly predicts expenses for those groups that are explicitly included in the model. Hence, to evaluate the statistical fit of an OLS-model on groups it is required to define groups that are *not* identical to those defined by the risk

adjusters (see Chapter 2). If, however, another statistical model specification than OLS is used it may be of interest to evaluate the predictive performance on the same groups as included in the model because then average predicted expenses do not necessarily have to equal average observed expenses for these groups.

§ 1.6 GOAL AND STRUCTURE OF THIS THESIS

The goal of this thesis is to provide insight into the evaluation of the predictive performance of RE-models (Part I) and to what extent the predictive performance of a morbidity-based RE-model can be improved by including additional risk adjusters based on prior years' costs and/or prior years' diagnostic information, or interaction terms between existing risk adjusters, or risk adjusters or interaction terms based on prior years' residual expenses (Part II)[8].

The remainder of this thesis is structured as follows. In *Part I*, Chapter 2 examines research question 1. Then, Chapter 3 elaborates on the preferred evaluation method as concluded in Chapter 2, aiming to provide an answer to question 2. The principles for evaluating models' predictive performance as formulated in these chapters are generally applicable to any RE-model.

In *Part II*, Chapters 4 through 6 each examine potentially relevant risk adjusters or interaction terms to improve model's predictive performance, aiming to provide answers to research questions 3 through 5, respectively. In each of these chapters, a comparative model evaluation is conducted to examine to what extent the performance of the benchmark model can be improved.

Chapter 7 summarizes the main findings from the preceding chapters, aiming to provide an answer to the central question of this thesis. In addition, this chapter provides reflections on the findings from preceding chapters and discusses the methods that have been used and the relevance of these findings for RE-models that are used around the world. Finally, Chapter 7 provides some recommendations and offers some directions for further research.

[8] This thesis is based on five publications. The first author of all publications has performed most of the work during all stages of the research, starting from searching relevant literature, to performing the empirical analyses, and to the final stage of reporting findings. The co-authors were consulted on a frequent basis to share ideas, discuss findings, and/or provide comments on the manuscript. For the empirical analysis, the same data as used in the Netherlands for calculating the actual RE-model were used. Data from the years 2006 to 2009 did not have to be collected anymore at the start of this thesis. The first author merged the data and performed the analyses on these data. Since 2013, the first author was a member of the research-team that worked on calculating the actual risk-adjusted payments in the Netherlands and so, she has contributed to the process of analyzing raw data to construct the administrative datasets from the years 2010 and 2011, which are used for the empirical analysis in this thesis as well.

Chapters 2 through 6 are written as separate publications and therefore, they can be read independently.

Part I

Evaluating the Predictive Performance of Risk Equalization Models

Chapter 2

A Taxonomy and Review of Measures-of-Fit

ABSTRACT

This chapter provides a taxonomy of measures-of-fit that have been used for evaluating risk equalization models since 2000 and discusses important properties of these measures, including variations in analytic method. It is important to consider the properties of measures-of-fit and variations in analytic method, because they influence the outcomes of evaluations that eventually serve as a basis for policymaking. Analysis of 81 eligible studies resulted in the identification of 71 unique measures that were divided into 3 categories based on treatment of the prediction error: measured based on squared errors, untransformed errors, and absolute errors. We conclude that no single measure-of-fit is best across situations. The choice of a measure depends on preferences about the treatment of the prediction error and the analytic method. If the objective is measuring financial incentives for risk selection, the only adequate evaluation method is to assess the predictive performance for non-random groups.

§ 2.1 INTRODUCTION

§ 2.1.1 Background

Risk equalization (RE) is a mechanism that provides health insurers or health plans with risk-adjusted payments to compensate them for differences in individuals' expected healthcare expenses. Over the past few decades, several countries worldwide have implemented RE in their health insurance systems, including Belgium, Germany, Israel, the Netherlands, Switzerland, and the U.S. (Medicare) (van de Ven et al., 2007; Pope et al., 2004). Recently, RE was introduced in the health insurance exchanges in the U.S. (Kautter et al., 2014). Given premium regulation in all of the aforementioned countries, RE aims to mitigate financial incentives for risk selection and thereby to achieve a level playing field for health insurers or health plans.

Over the past fifteen years, a vast amount of literature has focused on improving the predictive performance of RE-models. RE-models use several risk factors to predict individuals' healthcare expenses which are used as the basis for risk-adjusted payments[1]. Numerous empirical evaluations have been performed to assess and compare RE-models' predictive performance. Examples of such evaluations are studies examining (i) the effect of adding particular risk adjusters (e.g. Ash et al., 2005); (ii) the effect of using different data sources, such as demographic, pharmacy, or health survey information (e.g. Pietz, et al., 2004); or (iii) overcompensation or under-compensation for specific groups (e.g. Levy, et al., 2006). Though these evaluations differ in their study objective and method, they all have used one or more measures-of-fit to assess models' predictive performance.

§ 2.1.2 Prior research

Although many empirical evaluations have been performed in the RE literature, the properties of the measures-of-fit that were used have not been systematically studied. A few studies have attended to some properties of some measures and how they have influenced the outcomes of evaluations (e.g. Ash & Byrne-Logan, 1998; Ash et al., 2005; van de Ven & Ellis, 2000; Cumming et al., 2002). However, no comprehensive overview exists of the measures that are used and how they are applied.

An illustration of the importance of understanding the influence of the properties of the measures-of-fit and how they are applied in the outcomes of evaluations is the existence of

[1] The exact calculation of the risk-adjusted payment to health insurers or health plans differs across countries. The risk-adjusted payment can equal predicted expenses of an individual as calculated by the RE-model, e.g. this is the case in Israel, or it can equal the predicted expenses minus the average premium the individual pays to the health insurer out-of-pocket, e.g. this is the case in the Netherlands. These country-specific differences in modalities are not important for the objective of this study, since in all countries, regardless of modality, it is important that the RE-model adequately predicts individuals' expenses.

a number of misconceptions in the literature about some measures-of-fit. With misconceptions we refer to situations where measures have not been applied appropriately or situations where measures have been applied appropriately, but the outcomes are presented in such a way that they can lead to misinterpretation. An evident example is the comparison of R-squared (R^2)-values across studies, when there are differences in datasets, settings, and methods. Differences in R^2-values could be misinterpreted as differences in models' predictive performance, while they can be due to these other differences. Such misconceptions emphasize the need for a comprehensive overview of the measures-of-fit that are used, together with a critical assessment of their properties and variations in analytic method.

§ 2.2 NEW CONTRIBUTION

To our knowledge, this study is the first to provide a taxonomy of the measures-of-fit that have been used for assessing the predictive performance of RE-models and to discuss important properties of those measures, including variations in analytic method. We conducted a systematic literature review to obtain a comprehensive overview of the measures that have been used for evaluating RE-models since 2000.

Several studies show that even the most extended RE-models with morbidity-based risk adjusters do not predict expenses adequately for non-random groups (e.g. Behrend et al., 2007; van Kleef et al., 2014; Payne, et al., 2000; Pope et al., 2004). Empirical evaluations can help to improve existing RE-models and design RE-models for new types of expenses, such as long-term care services or mental care services. This study can contribute to these developments by informing researchers and policymakers about important properties of the measures-of-fit that can be used for evaluating RE-models and how to apply these measures appropriately. Several studies show that different measures and variations in applying the same measure can lead to different outcomes, which can lead eventually to different decisions concerning the design of the RE-model (e.g. Fishman et al., 2003; Powers et al., 2005). This study attempts to clarify some misconceptions about some measures-of-fit.

The next section describes how to place measures-of-fit in a broader context of evaluation measures and criteria for RE-models. This section also discusses the specification of RE-models. Thereafter, we outline the strategy for conducting the systematic literature review, followed by presenting the results of our search. Further, this section provides the taxonomy of the measures-of-fit and a detailed discussion of variations in analytic method. The final section discusses our findings.

§ 2.3 THEORETICAL FRAMEWORK

§ 2.3.1 Evaluation measures and criteria

The focus of this study is strictly on measures-of-fit; i.e. measures quantifying the extent to which a model predicts expenses for individuals or groups. This study does *not* focus on quantitative measures used for assessing the power and balance of payment schemes (Geruso & McGuire, 2014) or the statistical significance of risk adjusters, such as *T*-statistics. Besides models' predictive performance, there are also other, more qualitative, evaluation criteria, even though predictive performance has received the most attention in the RE literature. Examples of qualitative evaluation criteria are availability of data, appropriateness of incentives for risk selection and efficiency, and vulnerability to manipulation (van de Ven & Ellis, 2000). For policymakers, these qualitative criteria may be important in addition to models' predictive performance when deciding on the design of the RE-model.

§ 2.3.2 Specification of the statistical model

The RE-model is a statistical model that is used to predict individual expenses. At the estimation stage, the model is specified and validated and, if necessary, the model is re-specified and checks for goodness-of-fit and overfitting[2] are performed again. During this process, several measures can be used for model specification and validation, such as the Akaike Information Criterion, the Bayesian Information Criterion, and the Copas-test[3]. It is beyond the scope of this study to review measures that are used for model selection and validation[4]. This study focuses on measures that are used for assessing models' predictive performance, given the specification of the RE-model. Further, this study does not aim to discuss thoroughly how RE-models can accurately incorporate distributional properties of individual healthcare expenses[5]. For the purpose of this study, it is important to note that the specification of the statistical model restricts the choice of the measures-of-fit that can be used (see § 2.5). Researchers and policymakers should consider the specification of the RE-model when interpreting the outcomes of evaluations.

---

[2] Overfitting occurs when the model that has been estimated describes random error or noise instead of the underlying relationship in the data.

[3] This test is used for checking overfitting. In the literature, also referred to as 'Mincer-Zarnowitz'-test.

[4] Measures for model selection and validation are well-documented elsewhere: e.g. Fox, 2008; Hastie et al., 2008.

[5] For a thorough discussion on the development of statistical models for predicting individual healthcare expenses and how to incorporate the distributional properties of individual healthcare expenses see Duan et al., 1983; Manning & Mullahy, 2001; and Manning et al., 2005.

§ 2.4 METHOD

§ 2.4.1 Search strategy

We conducted systematic searches in four electronic databases: Medline (PubMed), ProQuest (e.g. Econlit), Scopus, and ISI web of knowledge. We used the same search terms, while taking into account the specific requirements of each database (see Appendix 2.1). Further, we screened the Embase and the Cochrane Library to find additional studies. These searches did not yield new relevant studies. We also screened the reference lists of the eligible studies and the publication lists of several authors well-known in the field of RE, and we consulted three experts to obtain additional relevant studies. Finally, we performed hand searches in Google and Google Scholar to check if we missed any important study[6].

§ 2.4.2 Selection process

The first author screened the databases and identified studies based on the eligibility criteria. During the selection process, the three other investigators were consulted multiple times to discuss findings. After the full-text screening by the first author, the three other investigators were consulted to judge the final list of selected studies against the eligibility criteria. Any uncertainty or disagreement was resolved through discussion.

§ 2.4.3 Eligibility Criteria

Table 2.1 presents the final set of eligibility criteria. Articles were eligible for inclusion when written in English and published in the period January 2000 - mid July 2013[7]. Our search was restricted to studies published since 2000 because the availability and quality of data have considerably improved since then. Consequently, new types of risk adjusters were developed and many studies performed more complete empirical evaluations in the sense that they assessed models' predictive performance on the sample level *and* subsample level (i.e. groups). We assume that there are no useful measures-of-fit in the RE literature before 2000 that have not been applied since then. Articles in press, comments, editorials, or conference papers were excluded from our search.

The initial search focused on articles in the field of RE or risk adjustment in the context of risk-adjusted payments, also referred to as capitation. The term 'risk adjustment' is also used in many other contexts, such as case-mix adjustments in measuring clinical outcomes, practice variation, or pay-for-performance schemes. On the one hand, a broad range of search terms was used to avoid exclusion of relevant studies. On the other hand, the search was restricted to some extent by combining 'risk adjustment' with 'capitation', 'health

[6] Information on the additional search strategy can be provided on request.

[7] The search in the four electronic databases was completed on 11 July 2013. We aimed to find all studies from 1 January 2000 until the day of completing the search.

Table 2.1: Eligibility criteria

	Inclusion	Exclusion
Language	English	Other languages
Publication date	January 2000 - mid-July 2013	Before January 2000
Publication status	Articles in peer-reviewed journals, studies that have not been published in peer-reviewed journals are not excluded beforehand	Articles in press, editorials, comments, letters, conference papers or presentations, news, front page/cover stories
Availability of study	Full-text of article is (freely) available	No (free) access to the full-text of the article
Context of the study	RE within the context of risk-adjusted (capitation) payments to health insurers	Risk adjustment within another context than RE, such as practice variation or pay-for-performance
Study objective	Estimating one or more RE-models and assessing models' predictive performance	Studies estimating one or more RE-models but with another objective than assessing models' predictive performance
Type of study	Empirical studies with quantitative analytic methods	Theoretical studies or empirical studies using solely qualitative analytic methods
Type of quantitative analysis method	Cell-based approach or regression analysis methods	Other methods than the cell-based approach and regression analysis (e.g. data mining algorithms, neural networks, or regression trees)
Target	Prediction of expenses	Prediction of utilization or healthcare need
Type of dependent variable	The variable is measured on a continuous scale	The variable is measured on a categorical or dichotomous scale

insurer(s)', or 'health plan(s)' in order to narrow the search results and avoid identifying articles about clinical outcomes, practice variation, or pay-for-performance. Note that our search strategy does not guarantee that articles on these topics were all excluded. For an article to be selected, the following four criteria were met. First, the article should conduct an empirical evaluation examining the prediction of healthcare expenses in the context of risk-adjusted payments to health insurers or health plans. We did not exclude studies examining risk adjustment within a broad context, meaning both within the context of risk-adjusted capitation payments and performance assessment (e.g. Ash & Ellis, 2012). Second, the empirical evaluation should consist of estimating one or more RE-models using cell-based or regression analysis, followed by an assessment of models' predictive performance. Third, the dependent variable should be healthcare expenses, measured on a continuous scale. Some studies have estimated several models, each with other dependent variables. These studies have not been excluded, but we extracted only the measures used to evaluate the models with healthcare expenses on a continuous scale as the dependent variable. Fourth, studies should include a statistical analysis for the purpose of that study and present the outcomes of their evaluation. We included studies not performing a statistical analysis, but using the results from another study that was not been identified by our search, such

as national reports, if the data and method were adequately reported (e.g. Ash et al., 2005; Buchner et al., 2013). An article was excluded if one of the inclusion criteria was not met.

§ 2.4.4 Data extraction

From each eligible study, the measures-of-fit were extracted, meaning that we extracted measures examining the *prediction error* of the RE-model. The prediction error is defined as the difference between predicted expenses by the RE-model and observed expenses or another reference point. Measures not examining the prediction error were not extracted, since those measures do not assess models' predictive performance, such as means or standard deviations[8,9].

In addition to the measures-of-fit, information on the study objective, data, and method were extracted. Specifically, we focused on variations in analytic method. This information was used to discuss how to apply the measures-of-fit appropriately to evaluate RE-models. We performed the procedure of data extraction twice to minimize the chance of missing important information[10]. Some authors were consulted by email when there was uncertainty about the precise study method or specific context.

§ 2.5 RESULTS

§ 2.5.1 Search results

A total of 1,852 citations were obtained from searching the four databases, which led to 1,140 potentially-relevant articles after deleting 712 duplicates among the databases (Figure 2.1). After screening the titles and abstracts, 191 articles remained. Four studies were added based on references, publication lists, and expert-opinions, resulting in 195 studies for detailed full-text screening. Based on the eligibility criteria, 81 studies were included in our analysis (see Appendix 2.2).

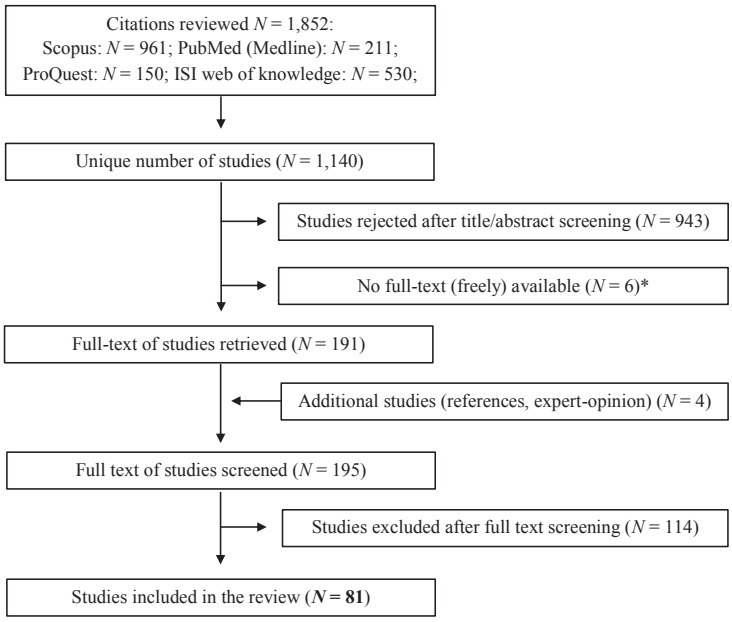
---

[8] Examples of excluded studies and examples of excluded measures can be provided on request.

[9] According to econometric theory, *prediction* errors should be calculated on a validation sample, which is an independent sample that is not used for model estimation (Hastie et al., 2008). We did not exclude studies using the full sample for prediction and evaluation, but indeed we aim to address the importance of using a validation sample for assessing models' predictive performance (see § 2.5.3).

[10] The process of data extraction has been performed twice by the first author. Ideally, this process would have been done by two investigators independently. However, due to the labor-intensity of analyzing all 81 studies in detail one investigator has done the entire process of data extraction. When there was uncertainty about the measure that was used or the study method, the other investigators were consulted. The information extracted from all studies has been reported in a separate document and can be provided on request.

Figure 2.1: Flow diagram for search and selection process



Footnote Figure 2.1:
* The full-text of these articles was not freely accessible via the sophisticated library of the Erasmus University Rotterdam. In the abstract of these studies we could not found any measure(s) of-fit of which we did not have knowledge about yet. Based on this, we conclude that exclusion of these studies did not influence our results about *which* measures have been applied since 2000.

§ 2.5.2 Taxonomy of the measures-of-fit

Based on our analysis of the eligible studies, 71 unique measures were identified. Most of these measures have great similarities in their underlying properties because the same measure has been applied in slightly different ways. To provide a taxonomy of these measures, as presented in Table 2.2, we used the following classification procedure. We reduced the 71 measures to 30 measures by aggregating across four important variations in analytic method: the level of analysis, the type of sample used for prediction and model evaluation, the reference point against which predicted expenses are compared, and standardization. Then, the 30 measures were further clustered to 6 measures on the basis of 3 methods of treating the prediction error: measures based on squared errors, untransformed errors, and absolute errors. More properties and variations in analytic method could be distinguished, but we believe that these are the most important to discuss. An overview of the 71 initial measures and detailed descriptions of these measures is provided in Appendix 2.3. The remainder of this section discusses the taxonomy of the measures in Table 2.2. The next section discusses the four distinguished variations in analytic method.

Table 2.2: Taxonomy of measures-of-fit that have been used for evaluating risk equalization models since 2000

Treatment of the prediction error	Measure	Measure variant (number of subtypes)	Study [a,b]
Squared error	R-squared (R^2)	conventional R^2 (5 subtypes)	1-6, 8-11, 13-16,18-22, 25-32, 34, 35, 39, 40, 42-51, 54-64, 66-71, 73-76, 78-81
		R^2 as a percentage of the maximum R^2 (3 subtypes)	6, 9, 10, 25, 37, 44-46, 57, 61
		Alternatives to the conventional R^2 (4 subtypes)	14, 17, 23, 37, 51
	Mean Squared Prediction Error (MSPE)	Conventional MSPE (3 subtypes)	51, 74, 75
		MSPE per decile of observed expenses (1 subtype)	51
		Root Mean Squared Prediction Error (RMSE) or RMSE to the mean (3 subtypes)	15, 54, 58
Untransformed errors	Predictive Ratio (PR) / Cost Ratio (CR)	(conventional) PR / CR (5 subtypes)	1-4, 7, 11, 13, 16, 18, 19, 22, 23, 25, 27-29, 31, 34, 35, 37, 39, 41, 43, 47, 51, 52, 54-58, 60, 62, 63, 65-67, 73, 79-81
		PR per decile or quintile or cost categories of observed or predicted expenses (3 subtypes)	9, 15, 25, 47, 49-51, 54, 55, 57, 66
		Predicted expenses to average observed expenses (E/A-ratio) (3 subtypes)	64, 65, 72
		Average Percentage Prediction Error Ratio (APPER) (2 subtypes)	42, 53, 72
		Mean Deviation Score (MDS) (1 subtype)	3
	Mean Prediction Error (MPE)	(conventional) MPE (6 subtypes)	5, 8, 9, 11, 12, 24, 25, 36, 44, 48, 68, 70, 71
		Average Percentage Prediction Error (APPE) (3 subtypes)	8, 68, 76, 77
		MPE per decile of predicted or observed expenses (2 subtypes)	25, 26
		Number and expenses for individuals with MPE below, within, or above X% of predicted expenses (1 subtype)	48
		MPE of one group divided by MPE for another group (2 subtypes)	33
		MPE for a group multiplied by the number of individuals within this group (1 subtype)	52

Table 2.2: (continued)

Treatment of the prediction error	Measure	Measure variant (number of subtypes)	Study [a,b]
Untransformed errors	Mean Prediction Error (MPE)	Total profit (1 subtype)	20
		Percentage of individuals with MPE greater or smaller than a certain percentage level (1 subtype)	24
		Range of average per-individual profit (1 subtype)	23
		Relative MPE (1 subtype)	36
		Percentage distribution of the error (2 subtypes)	38, 63
		Forecast bias (2 subtypes)	14, 74
Absolute errors	Mean Absolute Prediction Error (MAPE)	(conventional) MAPE (5 subtypes)	1, 3-6, 9, 11, 18, 23, 24, 25, 32, 40, 44, 56, 59, 62, 67, 68, 70, 71
		Mean Absolute Percentage Prediction Error (MAPPE) (4 subtypes)	1, 11, 15, 16
		MAPE per decile/quintile of actual expenses (1 subtype)	9, 25, 56
		Reduction in MAPE (1 subtype)	56
		'Standardized' MAPE (1 subtype)	30
		Mean Absolute Deviation (MAD) (1 subtype)	40
	Cumming's Prediction Measure (CPM) (2 subtypes)		3, 9, 11, 15, 18

Footnotes Table 2.2:

a. See reference list in Appendix 2.2 for the number that corresponds to each study.

b. For example, "1-6" means study 1, 2, 3, 4, 5, *and* 6 in the reference list.

Squared prediction errors: R-squared (R^2) and Mean Squared Prediction Error (MSPE)

Examples of measures based on squared prediction errors are the R^2, MSPE, and all variants of these measures. Measures based on squared errors weigh large errors more heavily than small errors, which make them sensitive to variance in expenses and outliers in the data.

Table 2.2 shows that the R^2 is the most commonly-used measure. The conventional R^2 can be calculated as one minus the ratio of the variance of the error divided by the variance of observed expenses[11]. The R^2-value ranges between zero and one, with a value closer to one indicating higher predictive performance. Several studies have calculated the R^2 as a percentage of the maximum R^2, of which some calculated the maximum R^2 by themselves (e.g. Breyer et al., 2003), while most others used the maximum R^2-value from another study (e.g. Fishman et al., 2003). Besides the conventional R^2 for linear models, the R^2 has also been used to evaluate nonlinear models, such as two-part models or generalized linear models. These types of R^2s have been calculated differently than those used for linear models. Within the RE literature, the Efron's R^2 (e.g. Kapur et al., 2000) and the generalized R^2 (Madden et al., 2000) have been used. These types of R^2s both have, as does the R^2 for linear models, a value ranging between zero and one, with a value closer to one indicating higher predictive performance.

The MSPE is the average of the squared differences between predicted and observed expenses over all individuals in the sample or subsample. Its value is always positive, with a smaller value indicating higher predictive performance. Besides the conventional MSPE, the MSPE has also been examined per decile of observed expenses (Madden et al., 2000). When comparing models, the model with the lowest MSPE-values across deciles has the highest predictive performance.

A measure closely related to the MSE is the 'Root Mean Squared Error' (RMSE), which is the square root of the MSE. Few studies have applied this measure (Pacala et al., 2003; Pope et al., 2004). Chalupka (2010) used the ratio of the RMSE to mean observed expenses. The RSME-value is always positive, with a smaller value indicating higher predictive performance.

Untransformed prediction errors: Predictive Ratio (PR) and Mean Prediction Error (MPE)

Examples of measures based on untransformed prediction errors are the PR, MPE, and all variants of these measures. These measures involve summary of untransformed errors

[11] The conventional R^2 can be calculated as: the variance of predicted expenses divided by the variance of observed expenses, or the square of the correlation between predicted expenses and observed expenses, or one minus the ratio of the variance of the error divided by the variance of observed expenses. These three ways of calculating the R^2 are only equivalent when an OLS-model is used and expenses are predicted on the full sample. If this is not the case, the only correct way of calculating the R^2 is the third method (Cameron & Windmeijer, 1996).

in which opposite signs of the errors can cancel each other out. An evident example is the evaluation of an Ordinary Least Squares (OLS)-model for the full sample (i.e. the same sample as used for model estimation). Because the sum of predicted expenses equals the sum of observed expenses in this situation, the PR will be one and the MPE will be zero, no matter which set of risk adjusters the RE-model uses; for example, a simple demographic model performs equally well as a model with morbidity-based risk adjusters. For this reason, these measures have not only been calculated for the full sample, but they also have been calculated for groups and for models applied on a validation sample (i.e. data not used for model estimation) or a reference point other than observed expenses is used to calculate the prediction error, which we explain in more detail below.

The PR is a commonly-used measure and is calculated as the ratio of predicted expenses to observed expenses. The terms 'Predicted-to-Observed', 'Cost Ratio', or 'Observed-to-Expected ratio' also have been used in this literature, with the latter two being reciprocals of the PR. The value of the PR is always positive, with a value closer to one indicating higher predictive performance. Besides the conventional PR, several studies have calculated the PR per decile or quintile of observed or predicted expenses (e.g. Levy et al., 2006).

Two other measures are the 'Predicted expenses to average observed ratio' (E/A-ratio) and the 'Average Percentage Prediction Error Ratio' (APPER). The E/A-ratio is the ratio of predicted expenses to average observed expenses of a reference population or group (e.g. Temkin-Greener et al., 2001). This measure indicates the proportion of higher-than-average observed expenses. Note that the E/A-ratio for the full sample equals the PR. The APPER, also referred to as the 'relative error' (Temkin-Greener et al., 2001) is calculated as follows. First, the ratio of observed expenses to average observed expenses (observed ratio) and the ratio of predicted expenses to average predicted expenses are calculated (predicted ratio). Then, the observed ratio is divided by the predicted ratio or vice versa, minus one multiplied by 100 percent (Noyes et al., 2006; Temkin-Greener et al., 2001). When evaluating an OLS-model for the full sample, the APPER-value for individuals or groups is the PR or the reciprocal of the PR minus one, expressed in percentages of observed or predicted expenses. Ash et al. (2005) refer to this measure as the 'Mean Deviation Score' and used it to indicate the percentage error of the prediction model. APPER-values closer to zero indicate higher predictive performance. Note that the APPER-value for the full sample equals zero, so this measure is only effective at the subsample level.

The MPE summarizes how well the model on average predicts expenses for a defined population or group. It is calculated as the mean of the difference between predicted and observed expenses over all individuals in the (sub)sample. In the literature, various terms have been used to refer to this measure, including 'mean error', 'mean result', 'profit or loss', 'under- or over-payment', 'selection profit', or 'gross profit'. The latter four measures interpret the MPE-value from the perspective of the health insurer; i.e. positive MPE-values indicate

profits or over-payments, while negative MPE-values indicate losses or under-payments. A MPE-value closer to zero indicates higher predictive performance.

Several alternatives to the MPE also have been used. Some studies have used the 'Average Percentage Prediction Error' (APPE), which is the MPE as a percentage of mean predicted expenses or mean observed expenses. When evaluating an OLS-model for the full sample, the APPE-value equals zero[12]. Other studies have examined the MPE per decile or quintile of observed or predicted expenses (e.g. Fishman et al., 2003).

Luft and Dudley (2004) used the MPE to examine the number of individuals and the expenses for those who have a MPE above a certain percentage of predicted expenses. Hsu et al. (2010) examined the MPE among different groups by calculating the MPE of one group to those of a reference group. Mark and colleagues (2003) did not divide the MPE of one group by another, but multiplied the MPE for each group by the number of individuals within this group in order to obtain the 'profit or loss' for each group.

Donato & Richardson (2006) used the MPE to examine the percentage of individuals with positive MPE-values in order to calculate the total profit for these individuals, while Ettner, and colleagues (2001) used the MPE to examine the percentage of individuals with a MPE-value above or below a certain critical percentage level, and Ettner and colleagues (2000) used the MPE to examine the range of average per-individual profit.

Kanters and colleagues (2013) used the 'relative MPE', which is based on the difference between the MPE-values of two models. Kautter and colleagues (2008) and Riley (2000) aimed to provide information about the distribution of the MPE. They split the full sample into several groups and calculated the MPE for each of them. Veazie and colleagues (2003) calculated the 'forecast bias', which is the difference between average predicted expenses and average observed expenses on a sample, averaged over multiple samples using random splitting.

Absolute prediction errors: Mean Absolute Prediction Error (MAPE) and Cumming's Prediction Measure (CPM)

Examples of measures based on absolute prediction errors are the MAPE, CPM, and all variants of these measures. These measures use the absolute value of the errors, meaning that they do not allow opposite signs of the error to cancel each other out, nor do they weigh errors differently. As such, these measures are less sensitive to extreme values in the distribution of expenses than measures based on squared errors (e.g. Buchner et al., 2013; Sales et al., 2003).

[12] In the literature, the APPE is also referred to as the 'percentage reduction in selection profit', 'percentage difference', 'average percentage under and overpayment', or 'gross profit rate'. The APPE is similar to the APPER, except that the APPE is expressed in deviations instead of ratios.

The MAPE is calculated as the mean of the absolute value of predicted expenses minus observed expenses across all individuals in the (sub)sample. The terms 'Mean Absolute Result' (e.g. Barneveld et al., 2000) and 'Mean Absolute Error' (Shen & Ellis, 2002) have also been used in this literature. The MAPE is expressed in absolute money amounts, with a value closer to zero indicating higher predictive performance. Several studies have used the 'Mean Absolute Percentage Prediction Error' (MAPPE), which is calculated as the MAPE as a percentage of mean observed expenses, averaged across all individuals (e.g. Chang & Weiner, 2010) or across health insurers (Buchner et al., 2013), or as a percentage of mean predicted expenses, averaged across several defined groups (Adams et al., 2002). Other studies have calculated the MAPE per decile of observed expenses (e.g. Fishman et al., 2003), while another study used the 'standardized' MAPE (Gilmer et al., 2001), which is the average of the absolute difference between standardized observed expenses and standardized predicted expenses for each group[13].

A measure closely related to the MAPE is the 'Mean Absolute Deviation' (MAD) (van Kleef & van Vliet, 2010), defined as the mean of the absolute difference between predicted and average observed expenses. This measure indicates the error of the model in predicting higher-than-average observed expenses, expressed in absolute terms.

Since the CPM was developed by Cumming and his colleagues (Cumming et al., 2002), it has been applied in only a few studies (e.g. Buchner et al., 2013; Behrend et al., 2007). The CPM is equal to one minus the ratio of the MAPE to the mean absolute difference between individual observed expenses and average observed expenses. Chalupka (2010) referred to this as the "proportion of explained sum of absolute errors". Like the R^2, its value ranges between zero and one, with a value closer to one indicating higher predictive performance.

Besides the three aforementioned methods of treating prediction errors, any measure in Table 2.2 can be transformed by explicitly weighting the errors. Van Barneveld and colleagues (2000) suggest ignoring small errors in order to account for transaction costs of health insurers and the statistical uncertainties about the net benefits of risk selection. To use this approach, policymakers need to decide how to weigh the prediction errors.

§ 2.5.3 Variations in Analytic Method

Level of analysis

RE-models can be evaluated at the sample level (i.e. the full sample over which the model is estimated), the subsample level (i.e. a part of the full sample/groups), or both. Almost all

[13] Standardized actual expenses have been calculated by dividing average actual expenses of each group by average actual expenses of a reference group. Standardized predicted expenses have been calculated by dividing average predicted expenses of each group by average predicted expenses of a reference group.

studies have evaluated models at the sample level, with many of them also at the subsample level. Evaluating models at the sample level may yield outcomes suggesting that the models predict reasonably well on average, while in fact the model can significantly under-predict or over-predict for non-random groups, which create financial incentives for risk selection. One method of identifying these under- and over-predictions is to perform an analysis on non-random groups. Analyzing the under- and over-predictions for random groups or insurers' portfolios is not adequate for measuring incentives for risk selection[14], because accurate predictions of expenses for random groups or insurers' portfolios can be the net effect of aggregation of significant under- and over-predictions for *non*-random groups. A reason why many studies have not assessed models' predictive performance for non-random groups may be lack of external data (i.e. data not used for estimation of the RE-model) to define these groups.

When evaluating models at the subsample level, it is important to consider which groups are defined. A model's measured performance is better when the groups are identical or closely related to the risk adjusters in the RE-model (van de Ven & Ellis, 2000; Cumming et al., 2002). For example, given an OLS-model, PRs for the same groups as the risk classes in the RE-model yield PRs equal or very close to one (e.g. Levy et al., 2006), which misrepresents model's true predictive performance. For this reason it is preferable to define groups that are not identical or closely related to the risk classes in the RE-model. For example, a health survey dataset that is not used for model estimation could be used to define the evaluation groups (e.g. Stam et al., 2010b). In some situations, however, external information is not available to define these groups. In these situations, the outcomes on groups that are similar or closely related to the risk classes in the RE-model should be interpreted with caution. In order to measure the extent to which an RE-model under- or over-predicts expenses for groups in the population, a representative sample is required. However, to measure under- and over-predictions for non-random groups, and thereby the existence of financial incentives for risk selection, it is *not* necessary to use a representative sample.

It is important to be aware of size and risk heterogeneity of the evaluation groups. By defining large groups, the uncertainty associated with random variation can be reduced. Outcomes on small groups can be more influenced by random variation than outcomes on large groups. Further, groups should be as homogeneous as possible, because a model can perform well on average for heterogeneous groups, while there can be significant under- or over-predictions for more homogeneous subgroups.

[14] When the objective of evaluating RE-models is measuring the level playing field, analysis at the portfolio level is appropriate. Hence, a level playing field does not guarantee absence of financial incentives for risk selection, because the outcomes of evaluations at the portfolio level also depend on the accidental risk composition of insurers' portfolio.

When performing a group-level analysis, a separate outcome is obtained for each group. When many groups are defined, it can be difficult to judge the overall performance of RE-models, because different models can perform differently on different groups. Interpreting the outcomes of a group-level analysis may require judging the relative importance of the prediction errors on different groups to decide whether one RE-model is preferred over another. This situation can be avoided by using a single-number summary measure describing the full sample. However, this type of analysis cannot provide information about the underlying performance of the RE-model for non-random groups.

Though in principle all measures in Table 2.2 can be applied at the sample level and subsample level, not all measures are meaningful at both levels. For example, the MPE and PR are only effective at the subsample level, but not at the sample level when measuring the performance of an OLS-model. When using the sample over which the OLS-model is fit, predicted expenses will sum to observed expenses, so the MPE will be zero and the PR will be one.

Type of sample used for prediction and model evaluation

Though most studies have assessed models' predictive performance on a validation sample, several studies have assessed models' predictive performance on the full sample (e.g. Hsu et al., 2010). The subtle but important difference between the use of the full sample or a validation sample to assess models' predictive performance has not always been recognized as such in the RE literature. A major concern with applying measures-of-fit on the full sample is that the outcomes may overstate models' true predictive performance because overfitting can occur (Ettner et al., 2000; Wooldridge, 2003; Hastie et al., 2009). For this reason, it is preferred to apply the measures-of-fit on a validation sample. In principle, all measures in Table 2.2 can be applied on a validation sample.

A prerequisite for applying measures on a validation sample is availability of a longitudinal dataset or a cross-sectional dataet with a sufficient size to split the sample. When applying split sampling methods, it is preferred to use random splitting to reduce selection bias in the outcomes. When the sample size is too small to split the sample and no longitudinal data are available, bootstrapping or K-fold cross-validation methods could be used to increase sample sizes (Ellis & Mookim, 2009; Hastie et al., 2009). Although in some situations these techniques can be useful, they have been applied rarely in the RE literature. It is worth noting that bootstrapped samples are not fully independent samples because the estimation sample and validation sample have many observations in common. For this reason, cross-validation techniques use non-overlapping data to construct a validation sample. Preferably, a validation sample is external to the estimation sample, but similar in terms of expenses and risk characteristics, so the measures-of-fit adequately reflect models' predictive performance.

Reference point for calculating the prediction error

The conventional method of calculating the prediction error of the measures in Table 2.2 is taking the difference between predicted expenses and observed expenses. There are some studies, however, that use a reference point other than observed expenses. One is normative expenses, defined as the expenses for which the policymaker deems compensation appropriate. Stam and his colleagues (2010a) predicted normative expenses using a prediction model with only risk adjusters for which the policymaker deems compensation appropriate. They used the normalized R^2, normalized MPE, and normalized MAR. Except for the reference point used, these measures have similar properties to their conventional versions. Another reference point used in the literature is expenses predicted by a model that includes more risk adjusters than the RE-model uses (Barneveld et al., 2000, 2001; Lamers, 2001). Stam and his colleagues (2010a) have advocated that normalized measures yield a better indication of models' true predictive performance compared to conventional measures because RE-predicted expenses cannot, and do not have to, equal observed expenses; e.g. it may not be desirable to adjust payments to health insurers for inefficiencies in the provision of services. Consequently, the outcomes of evaluations using measures with observed expenses as the reference cannot, and do not have to, equal the theoretical upper or lower bound of the measures, e.g. a R^2 of one or a MAPE of zero. Measures with a reference point other than observed expenses identify the appropriate norms. However, this method has not been widely adopted because these norms can be defined in many different ways and data are not available in many situations to apply this method. These are probably the reasons why few studies have used normalized measures.

Standardization

The measures in Table 2.2 can be divided into standardized and unstandardized measures. Examples of standardized measures are the R^2, PR, and CPM. Examples of unstandardized measures are the MSE, MPE, and MAPE. An advantage of standardized measures is that their value does not change when observed expenses or predicted expenses are expressed in another scale of measurement. With unstandardized measures, the value changes when observed or predicted expenses are expressed in another scale of measurement. For example, if observed expenses and predicted expenses are multiplied by a constant, the value of unstandardized measures will change, but the value of standardized measures remains the same. Consequently, when different scales of measurement are used, it is easier to use standardized measures than unstandardized measures for comparing performance across models. However, when it is relevant to interpret the differences in scale of measurement, for example, comparing the performance of one model on groups with different cost levels, unstandardized measures should be used. It is worth noting that standardization is not a distinctive measure property. Unstandardized measures can be standardized, for example, the MAPE as a percentage of mean observed expenses.

§ 2.6 CONCLUSIONS AND DISCUSSION

No comprehensive overview exists for measures-of-fit that are used for evaluating RE-models and how to apply these measures. This study conducted a systematic literature review to provide a taxonomy of measures-of-fit for evaluating RE-models and to discuss some important properties of these measures, including variations in analytic method. It is important to consider the properties of measures-of-fit and variations in analytic method because they influence the outcomes of evaluations that eventually serve as a basis for policymaking.

Analysis of 81 eligible studies resulted in the identification of 71 unique measures. Our taxonomy divides these measures into 3 categories based on treatment of the prediction error, with variations in analytic method. These 3 categories are: measures based on squared errors (e.g. R-squared (R^2) or Mean Squared Prediction Error (MSPE)); measures based on untransformed errors (e.g. Mean Prediction Error (MPE) or Predictive Ratio (PR)); and measures based on absolute errors (e.g. Mean Absolute Prediction Error (MAPE) or Cumming's Prediction Measure (CPM)). We examined four important variations in analytic method: the level of analysis, the type of sample used for prediction and model evaluation, the reference point against which predicted expenses are compared, and standardization.

Based on analysis of the properties of the measures and their variations in analytic method, we conclude that no one measure-of-fit is best across situations. The choice of a measure depends on preferences about the treatment of the prediction error and the analytic method. As several authors have advocated, there is no single measure that is comprehensive and sensitive in discriminating the predictive performance of different models (Ash et al., 2000; Cumming et al., 2002; Fishman et al., 2003).

Although there is no single measure-of-fit that is best across situations, there is only one appropriate analytic method if the objective of evaluating RE-models is measuring financial incentives for risk selection. To measure financial incentives for risk selection, the evaluation should include measurements of predictive performance for non-random groups that could be identified on the basis of observable characteristics. Analyses for the full sample, random groups, or insurers' portfolio level are not adequate for this purpose because the outcomes of these analyses can be the net effect of aggregation of significant under- and over-predictions for non-random groups.

To perform an analysis on groups, it is important to use external data (i.e. data not used for model estimation) to define groups that are not identical or closely related to the risk classes in the RE-model to prevent overfitting. Indeed, the outcomes of evaluations for groups that are identical or closely related to the risk classes in the RE-model and are defined on the same data as used for model estimation should be interpreted with caution because overfitting can occur.

When the evaluation of fit is conducted on non-random groups, several principles can be applied to decide on measures-of-fit. First, the MPE can be useful because this measure expresses the prediction errors for groups in money amounts. To apply the MPE to groups, it is important that homogeneous groups are specified because the MPE allows negative errors to cancel out positive errors. Second, when heterogeneous groups are defined, it can be useful to use a combination of the MPE and the MAPE. This is because the MPE-value can be zero, while there can be under- and over-predictions for homogeneous subgroups. The MAPE prevents negative errors from cancelling out positive errors. The MAPE, however, can only provide information about whether RE-models predict expenses adequately for groups, but not whether these models under- or over-predict expenses for these groups. Moreover, the lower bound of the MAPE that can be achieved with RE-models is unknown, while a MPE-value of zero indicates that the RE-model adequately predicts expenses on average for the defined group. Third, the R^2 and PR have often been used to evaluate RE-models; however, the use of these measures as indicators of fit on groups is limited. This is because the maximum R^2 that can be achieved with RE-models is unknown and the PR value does not describe the monetary value of the prediction error represented by a particular PR, which differs depending on the cost of a particular group. Presentation of the monetary value of the prediction error may be relevant to decide on the relative importance of accurate predictions for different groups. To conclude, because each measure has pros and cons, it is helpful to use multiple measures for evaluating the same RE-model.

When different measures are used, interpretation problems can occur when they produce conflicting outcomes. If there is no clear evidence that one model outperforms all others based on its predictive performance, evaluation criteria other than models' predictive performance alone, such as availability of data or appropriateness for incentives for risk selection and efficiency, can help to decide on the design of the RE-model. Moreover, when many evaluation groups are defined, it can occur that RE-models predict expenses better for some groups, but worse for others. In these situations, policymakers need to decide how to weigh the outcomes for different groups, which may involve using other evaluation criteria in addition to models' predictive performance.

§ 2.6.1 Study Limitations

This study is limited in that our analysis focused strictly on measures used for assessing the predictive performance of RE-models. Decisions about the design of the RE-model are likely to depend on more criteria than solely predictive performance. Policymakers may weigh the outcomes of evaluations of RE-models against other considerations, such as the availability of data and appropriateness of incentives for risk selection and efficiency. In policymaking, these other evaluation criteria may be important in addition to models' predictive performance.

This study is also restricted to the retrieved studies and the information reported in those studies. Further, this study is limited to those measures-of-fit that have been used for evaluating RE-models since 2000. However, we screened handbooks in the financial and econometric literature and these searches did not yield measures that have not been used within the RE field (Wooldridge, 2003; Fox, 2008; Hastie et al., 2009).

Although this study focused on the RE literature, prediction models have been used in a much broader context; for example, in financial economics or actuarial sciences. This study did not include literature in these other scientific fields.

Appendices
Chapter 2

APPENDIX 2.1: SEARCH STRATEGY

General remark on the search terms: the term "risk assessment" has not been used as a keyword, since using this keyword resulted in studies on assessing the association of individuals' characteristics with expected use of medical services rather than risk equalization. The dependent variable in these studies was an indicator of medical healthcare use, instead of medical healthcare expenses.

Pubmed (*N* = 211; search on 11.07.2013)

"risk equalization"[Title/Abstract] OR ("risk adjustment"[Title/Abstract] AND "capitation"[All Fields]) OR ("risk adjustment"[Title/Abstract] AND "capitated payments"[All Fields]) OR ("risk adjustment"[Title/Abstract] AND "health plans"[All Fields]) OR ("risk adjustment"[Title/Abstract] AND "health plan"[All Fields]) OR ("risk adjustment"[Title/Abstract] AND "insurers"[All Fields]) OR ("risk adjustment"[Title/Abstract] AND "insurer"[All Fields]) OR "risk adjusted payments"[Title/Abstract] OR "capitation formula"[Title/Abstract] OR ("prediction model"[Title/Abstract] AND "expenses"[All Fields]) OR ("prediction model"[Title/Abstract] AND "expenses"[All Fields]) OR "expenses model"[Title/Abstract] OR "predicting expenses"[Title/Abstract] OR ("predictive accuracy"[Title/Abstract] AND "expenses"[Title/Abstract]) OR ("predictive accuracy"[Title/Abstract] AND "expenses"[Title/Abstract]) OR ("predictability"[Title/Abstract] AND "expenses"[Title/Abstract]) OR ("predictability"[Title/Abstract] AND "expenses"[Title/Abstract]) AND (("2000/01/01"[PDAT] : "2013/07/11"[PDAT]) AND English[lang])

Limits: English; Published from 01/01/2000 – 11/07/2013; (no option found for peer-reviewed articles)

ProQuest (*N* = 150*); search on 11.07.2013)

((((AB("risk equalization")) OR (AB("risk adjustment") AND FT(capitation)) OR (AB("risk adjustment") AND FT("capitated payments")) OR (AB("risk adjustment") AND FT(plans)) OR (AB("risk adjustment") AND FT(plan)) OR (AB("risk adjustment") AND FT(insurers)) OR (AB("risk adjustment") AND FT(insurer)) OR (AB("risk adjusted payments")) OR (AB("capitation formula")) OR (AB("prediction model") AND FT(expenses)) OR (AB("prediction model") AND FT(expenses)) OR (AB("expenses model")) OR (AB("predicting expenses")) OR (AB("predicting expenses")) OR (AB("predictive accuracy") AND AB(expenses)) OR (AB("predictive accuracy") AND AB(expenses)) OR (AB(predictability) AND AB(expenses)) OR (AB(predictability) AND AB(expenses))) AND peer(yes)) NOT (at.exact("Commentary" OR "Editorial" OR "News" OR "Conference" OR "Front Page/Cover Story") AND la.exact("ENG") AND pd(20000101-20130711)))

Limits: Peer-reviewed; English abstracts; Published from 01/01/2000 – 11/07/2013

* Footnote: The database found 202 results, but there were duplicate citations. In total, there were 150 unique results.

Scopus (*N* = 961**; search on 11.07.2013)

(TITLE-ABS-KEY("risk equalization")) OR (TITLE-ABS-KEY("risk adjustment") AND ALL(capitation)) OR (TITLE-ABS-KEY("risk adjustment") AND ALL("capitated payments")) OR (TITLE-ABS-KEY("risk adjustment") AND ALL(plans)) OR (TITLE-ABS-KEY("risk adjustment") AND ALL(plan)) OR (TITLE-ABS-KEY("risk adjustment") AND ALL(insurers)) OR (TITLE-ABS-KEY("risk adjustment") AND ALL(insurer)) OR (TITLE-ABS-KEY("risk adjusted payments")) OR (TITLE-ABS-KEY("capitation formula")) OR (TITLE-ABS-KEY("prediction model") AND ALL(expenses)) OR (TITLE-ABS-KEY("prediction model") AND ALL(expenses)) OR (TITLE-ABS-KEY("expenses model")) OR (TITLE-ABS-KEY("predicting expenses")) OR (TITLE-ABS-KEY("predicting expenses")) OR (TITLE-ABS-KEY("predictive accuracy") AND TITLE-ABS-KEY(expenses)) OR (TITLE-ABS-KEY("predictive accuracy") AND TITLE-ABS-KEY(expenses)) OR (TITLE-ABS-KEY(predictability) AND TITLE-ABS-KEY(expenses)) OR (TITLE-ABS-KEY(predictability) AND TITLE-ABS-KEY(expenses)) AND (LIMIT-TO(PUBYEAR, 2013) OR LIMIT-TO(PUBYEAR, 2012) OR LIMIT-TO(PUBYEAR, 2011) OR LIMIT-TO(PUBYEAR, 2010) OR LIMIT-TO(PUBYEAR, 2009) OR LIMIT-TO(PUBYEAR, 2008) OR LIMIT-TO(PUBYEAR, 2007) OR LIMIT-TO(PUBYEAR, 2006) OR LIMIT-TO(PUBYEAR, 2005) OR LIMIT-TO(PUBYEAR, 2004) OR LIMIT-TO(PUBYEAR, 2003) OR LIMIT-TO(PUBYEAR, 2002) OR LIMIT-TO(PUBYEAR, 2001) OR LIMIT-TO(PUBYEAR, 2000)) AND (LIMIT-TO(DOCTYPE, "ar") OR LIMIT-TO(DOCTYPE, "re") OR LIMIT-TO(DOCTYPE, "ip")) AND (LIMIT-TO(LANGUAGE, "English"))

Limits: Publication data: 2000 – 2013; Type document: article, review; English

** Footnote: The database found 976 results, of which 12 articles were in press (meaning that they have not been published yet). These 12 articles were excluded. Further, there were 3 duplicates. In total, 961 results were obtained (976-12-3=961).

ISI web of Knowledge (*N* = 530***; search on 11.07.2013)

Topic=(("risk equalization")) OR Topic=(("risk adjustment" AND "capitation")) OR Topic=(("risk adjustment" AND "capitated payments")) OR Topic=(("risk adjustment" AND "plans")) OR Topic=(("risk adjustment" AND "plan")) OR Topic=(("risk adjustment" AND "insurer")) OR Topic=(("risk adjustment" AND "insurers")) OR Topic=(("risk adjusted payments")) OR Topic=(("capitation formula")) OR Topic=(("prediction model" AND "expenses")) OR Topic=(("prediction model" AND "expenses")) OR Topic=(("expenses model")) OR Topic=(("predicting expenses")) OR Topic=(("predicting expenses")) OR Topic=(("predictive accuracy" AND "expenses")) OR Topic=(("predictive accuracy" AND "expenses")) OR Topic=(("predictability" AND "expenses")) OR Topic=(("predictability" AND "expenses"))
Refined by: Document Types=(ARTICLE OR REVIEW) AND Publication Years=(2008 OR 2003 OR 2005 OR 2000 OR 2009 OR 2001 OR 2011 OR 2004 OR 2007 OR 2013 OR 2006 OR 2010 OR 2002 OR 2012) AND Languages=(ENGLISH)
Timespan=2000-2013.

Limits: Publication data: 2000 – 2013; Type document: article, review; English

*** Footnote: The database found 535 results, but there were 5 duplicates. In total, 530 (=535-5) results were obtained.

APPENDIX 2.2: REFERENCES OF THE STUDIES INCLUDED IN OUR REVIEW

1. 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.
2. Ash, A.S., Ellis R.P., Pope, G.C., Ayanian, J.Z., Bates, D.W., Burstin, H., Iezzoni, L.I., MacKay, E., Yu, W. (2000). Using Diagnosis to describe populations and predict expenses. *Health Care Financing Review, 21*(3), 7-28.
3. Ash, A.S., McCall, N., Fonda, J., Hanchate, A., Speckman, J. (2005). Risk assessment of Military Populations to predict health care expenses and utilization. Final report: Research Triangle Institute, Washington, D.C. and Boston University School of Medicine, Boston.
4. Ash, A.S., Ellis, R.P. (2012). Risk-adjusted payment and performance assessment for primary care. *Medical Care, 50*(8), 643-653.
5. van Barneveld, E.M., Lamers, L.M., van Vliet, R.C.J.A., van de Ven, W.P.M.M. (2000). Ignoring small predictable profits and losses: A new approach for measuring incentives for cream skimming. *Health Care Management Science, 3*, 131-140.
6. van Barneveld, E.M., Lamers, L.M., van Vliet, R.C.J.A., van de Ven, W.P.M.M. (2001). Risk sharing as a supplement to imperfect capitation: a trade-off between selection and efficiency. *Journal of Health Economics, 20*, 147-168.
7. Barry, C.L., Weiner, J.P., Lemke, K., Busch, S.H. (2012). Risk adjustment in health insurance exchanges for individuals with mental illness. *American Journal of Psychiatry, 169*(7), 704-709.
8. Beck, K. (2000). Growing importance of capitation in Switzerland. *Health Care Management Science, 3*(2), 111-119.
9. 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.
10. Breyer, F., Heineck, M., Lorenz, N. (2003). Determinants of health care utilization by German sickness fund members - with application to risk adjustment. *Health Economics, 12*(5), 367-376.
11. Buchner, F., Goepffarth, D., Wasem, J. (2013). The new risk adjustment formula in Germany: implementation and first experiences. *Health Policy*, *109*, 253-262.
12. Buntin, M.B., Garber, A.M., McClellan, M., Newhouse, J.P. (2004). The expenses of decedents in the Medicare program: implications for payments to Medicare+Choice plans. *Health Services Research, 39*(1), 111-130.
13. Calderón-Larrañaga, A., Abrams, C., Poblador-Plou, B., Weiner, J.P., Prados-Torres, A. (2010). Applying diagnosis and pharmacy-based risk models to predict pharmacy use in Aragon, Spain: the impact of a local calibration. *BMC Health Services Research, 10*(22).
14. Carter, G.M., Bell, R.M., Dubois, R.W., Goldberg, G.A, Keeler, E.B., McAlearney, J.S., Post, E.P., Rumpel, J.D. (2000). A clinically detailed risk information system for cost. *Health Care Financing Review, 21*(3), 65-91.
15. Chalupka, R. (2010). Improving risk adjustment in the Czech Republic. Prague Economic Papers, 3, 236-250.
16. Chang, H., Weiner, J.P. (2010). An in-depth assessment of a diagnosis-based risk adjustment model based on national health insurance claims: the application of the Johns Hopkins Adjusted Clinical Group case-mix system in Taiwan. *BMC Medicine, 8*(7), 1-13.
17. Chang, R., Lin, W., Hsieh, C., Chiang, T. (2002). Healthcare utilization patterns and risk adjustment under Taiwan's national health insurance system. *Journal of the Formosan Medical Association, 101*(1), 52-59.

18. Cumming, R.B., Knutson, D., Cameron, B.A., Derrick, B. (2002). A comparative analysis of claims-based methods of health risk assessment for commercial populations. Society of Actuaries, USA.
19. DeSalvo, K.B., Jones, T.M., Peabody, J., McDonald, J., Fihn, S., Fan, V., He, J., Muntner, P. (2009). Health care expenses prediction with a single item, self-rated health measure. *Medical Care, 47*(4), 440-447.
20. Donato, R., Richardson, J. (2006). Diagnosis-based risk adjustment and Australian health system policy. *Australian Health Review, 30*(1), 83-99.
21. Duckett, S.J., Agius, P.A. (2002). Performance of diagnosis-based risk adjustment measures in a population of sick Australians. *Australian and New Zealand Journal of Public Health, 26*(6), 500-507.
22. Dudley, R.A., Medlin, C.A., Hammann, L.B., Cisternas, M.G., Brand, R., Rennie, D.J., Luft, H.S. (2003). The best of both worlds? Potential of hybrid prospective/concurrent risk adjustment. Medical Care, 41(1), 56-69.
23. Ettner, S.L., Frank, R.G., Mark, T., Smith, M.W. (2000). Risk adjustment of capitation payments to behavioral health care carve-outs: how well do existing methodologies account for psychiatric disability? *Health Care Management Science, 3*(2), 159-169.
24. Ettner, S.L., Frank, R.G., McGuire, T.G., Hermann, R.C. (2001). Risk adjustment alternatives in paying for behavioral health care under Medicaid. *Health Services Research, 36*(4), 793-811.
25. 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.
26. Fleishman, J.A., Cohen, J.W., Manning, W.G., Kosinski, M. (2006). Using the SF-12 health status measure to improve predictions of medical expenses. *Medical Care, 44*(5 suppl.), I-54-I-63.
27. Frogner, B.K., Anderson, G.F., Cohen, R.A., Abrams, C. (2011). Incorporating new research into Medicare risk adjustment. *Medical Care, 49*(3), 295-300.
28. García-Goñi, M., Ibern, P. (2008). Predictability of drug expenses: an application using morbidity data. *Health Economics, 17*(1), 119-126.
29. García-Goñi, M., Ibern, P., Inoriza, J.M. (2009). Hybrid risk adjustment for pharmaceutical benefits. *European Journal of Health Economics, 10*(3), 299-308.
30. Gilmer, T., Kronick, R., Fishman, P., Ganiats, T.G. (2001). The Medicaid Rx model: pharmacy-based risk adjustment for public programs. *Medical Care, 39*(11), 1188-1202.
31. Hanley, G.E., Morgan, S., Reid, R.J. (2010). Explaining prescription drug use and expenses using the Adjusted Clinical Groups case-mix system in the population of British Columbia, Canada. *Medical Care, 48*(5), 402-408.
32. Hsu, J., Huang, J., Fung, V. Price, M., Brand, R., Hui, R., Fireman, B., Dow, W.H., Bertko, J., Newhouse, J.P. (2009). Distributing $800 billion: an early assessment of Medicare part D risk adjustment. *Health Affairs, 28*(1), 215-225.
33. Hsu, J., Fung, V., Huang, J., Price, M., Brand, R., Hui, R., Fireman, B., Dow, W.H., Bertko, J., Newhouse, J.P. (2010). Fixing flaws in Medicare drug coverage that prompts insurers to avoid low-income patients. *Health Affairs, 29*(12), 2335-2342.
34. 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.
35. Hwang, W., Ireys, H.T., Anderson, G.F. (2001). Comparison of risk adjusters for Medicaid-enrolled children with and without chronic health conditions. *Ambulatory Pediatrics, 1*(4), 217-224.
36. Kanters, T.A., Brouwer, W.B.F., van Vliet, R.C.J.A., van Baal, P.H.M., Polder, J.J. (2013). A new prevention paradox: the trade-off between reducing incentives for risk selection and increasing the incentives for prevention for health insurers. *Social Sciences & Medicine*, *76*, 150-158.

37. Kapur, K., Young, A.S., Murata, D. (2000). Risk adjustment for high utilizers of public mental health care. *The Journal of Mental Health Policy and Economics, 3*(3), 129-137.
38. Kautter, J., Ingber, M., Pope, G.C. (2008). Medicare risk adjustment for the frail elderly. *Health Care Financing Review, 30*(2), 83-93.
39. Kautter, J., Ingber, M., Pope, G.C., Freeman, S. (2012). Improvements in Medicare part D risk adjustment: beneficiary access and payment accuracy. *Medical Care, 50*(12), 1102-1108.
40. van Kleef, R.C., van Vliet, R.C.J.A. (2010). Prior use of durable medical equipment as a risk adjuster for health-based capitation. *Inquiry, 47*(4), 343-358.
41. van Kleef, R.C., van Vliet, R.C.J.A. (2012). Improving risk equalization using multiple-year high cost as a health indicator. *Medical Care, 50*(2), 140-144.
42. 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.
43. Kuhlthau, K., Ferris, T.G., Davis, R.B., Perrin, J.M., Iezzoni, L.I. (2005). Pharmacy- and diagnosis-based risk adjustment for children with Medicaid. *Medical Care, 43*(11), 1155-1159.
44. Lamers, L.M. (2001). Health-based risk adjustment: Is inpatient and outpatient diagnostic information sufficient? *Inquiry, 38*(4), 423-431.
45. 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.
46. Lamers, L.M., van Vliet, R.C.J.A. (2004). The pharmacy-based cost group model: validating and adjusting the classification of medications for chronic conditions to the Dutch situation. *Health Policy, 68*(1), 113-121.
47. Levy, J.M., Robst, J., Ingber, M.J. (2006). Risk-adjustment system for the Medicare capitated ESRD program. *Health Care Financing Review, 27*(4), 53-69.
48. Luft, H.S., Dudley, R.A. (2004). Assessing risk-adjustment approaches under non-random selection. *Inquiry, 41*(2), 203-217.
49. Maciejewski, M.L., Liu, C.–F., Derleth, A., McDonell, M., Anderson, S., Fihn, S.D. (2005). The performance of administrative and self-reported measures for risk adjustment of veterans affairs expenses. *Health Services Research, 40*(3), 887-904.
50. Maciejewski, M.L., Liu, C.–F., Fihn, S.D. (2009). Performance of comorbidity, risk adjustment, and functional status measures in expenses prediction for patients with diabetes. *Diabetes Care, 32*(1), 75-80.
51. Madden, C.W., Mackay, B.P., Skillman, S.M., Ciol, M., Diehr, P.K. (2000). Risk adjusting capitation: applications in employed and disabled populations. *Health Care Management Science, 3*(2), 101-109.
52. Mark, T.L., Ozminkowski, R.J., Kirk, A., Ettner, S.L., Drabek, J. (2003). Risk adjustment for people with chronic conditions in private sector health plans. *Medical Decision Making, 23*(5), 397-405.
53. Noyes, K., Liu, H., Temkin-Greener, H. (2006). Cost of caring for Medicare beneficiaries with Parkinson's disease: Impact of the CMS-HCC risk-adjustment model. *Disease Management, 9*(6), 339-348.
54. Pacala, J.T., Boult, C., Urdangarin, C., McCaffrey, D. (2003). Using self-reported data to predict expenses for the health care of older people. *Journal of the American Geriatrics Society, 51*(5), 609-614.
55. 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.
56. Pietz, K., Ashton, C.M., McDonell, M., Wray, N.P. (2004). Predicting healthcare expenses in a population of veterans affairs beneficiaries using diagnosis-based risk adjustment and self-reported health status. *Medical Care, 42*(10), 1027-1035.

57. 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. (2000a). Principal inpatient diagnostic cost group model for Medicare risk adjustment. *Health Care Financing Review, 21*(3), 93-118.
58. 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. (2004). Risk adjustment of Medicare capitation payments using the CMS-HCC model. *Health Care Financing Review, 25*(4), 119-141.
59. Powers, C.A., Meyer, C.M., Roebuck, M.C., Vaziri, B. (2005). Predictive modeling of total healthcare expenses using pharmacy claims data: A comparison of alternative econometric cost modeling techniques. *Medical Care, 43*(11), 1065-1072.
60. Prinsze, F.J., van Vliet, R.C.J.A. (2007). Health-based risk adjustment: Improving the pharmacy-based cost group model by adding diagnostic cost groups. *Inquiry, 44*(4), 469-480.
61. Reid, R.J., MacWilliam, L., Verhulst, L., Roos, N., Atkinson, M. (2001). Performance of the ACG case-mix system in two Canadian provinces. *Medical Care, 39*(1), 86-96.
62. Rein, D.B. (2005). A matter of classes: stratifying health care populations to produce better estimates of inpatient expenses. *Health Services Research, 40*(4), 1217-1233.
63. Riley, G.F. (2000). Risk adjustment for health plans disproportionately enrolling frail Medicare beneficiaries. *Health Care Financing Review, 21*(3), 135-148.
64. Robinson, J., Karon, S.L. (2000). Modeling Medicare expenses of PACE populations. *Health Care Financing Review, 21*(3), 149-170.
65. Robst, J. (2009). Development of a Medicaid behavioral health case-mix model. *Evaluation Review, 33*(6), 519-538.
66. Robst, J., Levy, J.M., Ingber, M.J (2007). Diagnosis-based risk adjustment for Medicare prescription drug plan payments. *Health Care Financing Review, 28*(4), 15-30.
67. Sales, A.E., Liu, C.F, Sloan, K.L., Malkin, J., Fishman, P.A., Rosen, A.K., Loveland, S., Paul Nichol, W., Suzuki, N.T., Perrin, E., Sharp, N.D., Todd-Stenberg, J. (2003). Predicting expenses of care using a pharmacy-based measure risk adjustment in a veteran population. *Medical Care, 41*(6), 753-760.
68. Shen, Y., Ellis, R.P. (2002). How profitable is risk selection? A comparison of four risk adjustment models. *Health Economics, 11*(2), 165-174.
69. Shmueli, A., Messika, D., Zmora, I., Oberman, B. (2010). Health care expenses during the last 12 months of life in Israel: estimation and implications for risk-adjustment. *International Journal of Health Care Finance and Economics, 10*(3), 257-273.
70. Stam, P.J.A., van Vliet, R.C.J.A., van de Ven, W.P.M.M. (2010a). A limited-sample benchmark approach to assess and improve the performance of risk equalization models. *Journal of Health Economics, 29*, 426-437.
71. Stam, P.J.A., van Vliet, R.C.J.A., van de Ven, W.P.M.M. (2010b). Diagnostic, pharmacy-based, and self-reported health measures in risk equalization models. *Medical Care, 48*(5), 448-457.
72. Temkin-Greener, H., Meiners, M.R., Gruenberg, L. (2001). PACE and the Medicare+Choice risk-adjusted payment model. *Inquiry, 38*(1), 60-72.
73. Vargas, V., Wasem, J. (2006). Risk adjustment and primary health care in Chile. *Croatian Medical Journal, 47*(3), 459-468.
74. Veazie, P.J., Manning, W.G., Kane, R.L. (2003). Improving risk adjustment for Medicare capitated reimbursement using nonlinear models. *Medical Care, 41*(6), 741-752.
75. Vivas, D., Guadalajara, N., Barrachina, I., Trillo, J.–L, Usó, R., De-La-Poza, E. (2011). Explaining primary healthcare pharmacy expenses using classification of medications for chronic conditions. *Health Policy, 103*(1), 9-15.

76. van Vliet, R.C.J.A. (2006). Free choice of health plan combined with risk-adjusted capitation payments: are switchers and new enrollees good risks? *Health Economics, 15*(8), 763-774.
77. Weiner, J.P., Trish, E., Abrams, C., Lemke, K. (2012). Adjusting for risk selection in state health insurance exchanges will be critically important and feasible, but not easy. *Health Affairs, 31*(2), 306-315.
78. Wrobel, M.V., Doshi, J., Stuart, B.C., Briesacher, B. (2004). Predictability of prescription drug expenses for Medicare beneficiaries. *Health Care Financing Review, 25*(2), 37-46.
79. Yu, H., Dick, A.W. (2010). Risk-adjusted capitation rates for children: how useful are the survey-based measures? *Health Services Research, 45*(6, part 2), 1948-1962.
80. Yuen, E.J., Louis, D.Z., Loreto, P.D., Gonnella, J.S. (2003). Modeling risk-adjusted capitation rates for Umbria, Italy. *European Journal of Health Economics, 4*(4), 304-312.
81. Zhao, Y., Ash, A.S., Ellis, R.P., Ayanian, J.Z., Pope, G.C., Bowen, B., Weyuker, L. (2005). Predicting pharmacy expenses and other medical expenses using diagnoses and drug claims. *Medical Care, 43*(1), 34-43.

APPENDIX 2.3: DESCRIPTION OF THE MEASURES-OF-FIT

This appendix – Table A.2.1 – lists how all measures that have been used for assessing the predictive performance of RE-models since 2000 have been classified, resulting into Table 2.2 in the main text. The measures have been clustered based on similarities and differences in their treatment of the prediction error and analytic method. The most important variations in analytic method are discussed in the main text. A general remark on Table A.2.1 is that this table is *not* exhausting. In theory there are more alternatives to the measures than those reported in this table. We only report those measures that have been used in the RE literature since 2000 up and until July 2013.

Table A.2.1: Description of how the measures-of-fit that have been used for evaluating the predictive performance of risk equalization models since 2000 have been clustered

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
1. (conventional) R^2	The R^2 is a goodness-of-fit measure. It is a measure for the proportion of variance in expenses explained by the model. The R^2 value can be adjusted for the number of variables in the model; the so-called adjusted R^2-value. We did not distinguish the unadjusted R^2 from the adjusted R^2, because it is a 'minor' property. Many studies did not report whether they applied the adjusted or unadjusted R^2, because the penalty for the number of included variables is not relevant anymore when a (very) large datasets is used for model estimation.	- Sample level; - In-sample; - Observed expenses; - Relative measure;	**1. Estimated R^2**	1, 2, 4-6, 8, 10, 11, 13-15, 18, 20-22, 25-32, 34, 35, 39, 40, 42-47, 49-51, 54, 55-64, 66-68, 71, 73-76, 78-81
		- Sample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**2. Validated R^2**	2, 3, 8, 9, 14, 16, 18-20, 25, 28, 31, 34, 42, 56, 59, 62, 67, 68, 71
		- Subsample level; - In-sample; - Observed expenses as reference; - Relative measure;	**3. Estimated Grouped-R^2**	4, 31, 69, 78
		- Subsample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**4. Validated Grouped-R^2**	3, 16, 25, 48, 62
		- Sample level; - In-sample; - Other reference than observed expenses; - Relative measure;	**5. Normalized R^2**	70

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
2. R^2 as a percentage of the maximum R^2	This measure is used to indicate to potential to which the R^2 could be improved. The R^2-value has an upper bound of one, but due to the unpredictability of healthcare expenses this theoretical upper-bound cannot be achieved with a prediction model. Instead of this theoretical upper bound, the maximum R^2 is used to indicate the maximum predictable variance in individual healthcare expenses. The maximum R^2 can be calculated in the study itself (this is the preferred method), or the max R^2 computed in another study can be used (not the preferred method). The maximum R^2 can be calculated by estimating an error components model (e.g. a variance-components model, an autoregressive model, a mixed autoregressive-variance components model, or an autoregressive-moving average model). The autocorrelations of the individual prediction errors are used to estimate the maximum proportion of variance in the error term that is predictable. It is based on the principle that the error of an individual in a certain year yields information on future expenses. See for exact calculations Newhouse et al. 1989, van Vliet 1992, Breyer et al. 2003, or Kapur et al. 2000.	- Sample level; - In-sample; - Observed expenses as reference; - Relative measure;	**6. Estimated R^2 as a % of max R^2**	6, 44[c]-46[c], 57, 61 Remark: Estimated R^2, with max R^2 calculated in another study has been used.
		- Sample level; - In-sample; - Observed expenses as reference; - Relative measure	**7. Estimated R^2 as a % of max R^2**	10, 37 Remarks: The estimated (conventional) R^2; the max R^2 calculated in the study is used
		- Sample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**8. Validated R^2 as a % of max R^2**	9, 25 Remarks: Validated (conventional) R^2; Max R^2 calculated another study has been used.

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
3. Alternatives to conventional R^2	Besides the conventional R^2, there are some other measures closely-related to the R^2. One of them is Efrons' R^2. Efrons' R^2 is also a goodness-of-fit measure. This measure, also called the synthetic R^2, indicates the proportion explained variance in expenses when the model is non-linear or when an out-of-sample is used to predict expenses. Another alternative is the Generalized R^2, which is also used for non-linear models.	- Sample level; - In-sample; - Observed expenses as reference; - Relative measure;	**9. (Estimated) Efron's R^2 / synthetic R^2**	37
		- Sample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**10. (Validated) Efron's R^2 / synthetic R^2**	14, 17, 23, 37
		- Subsample level; - Out-of-sample; - Observed expenses as a reference; - Relative measure;	**11. (Validated) Efron's (grouped) R^2 / synthetic R^2**	14
		- Sample level; - In-sample; - Observed expenses as a reference; - Relative measure; *Another statistical model than OLS.	**12. Generalized R^2**	51

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
4. (conventional) MSPE (Mean Squared Prediction Error / Mean squared error / Forecast mean squared forecast error)	A MSPE is calculated as the average of the squared prediction errors. Like the R^2, it is sensitive to outliers, because of the quadratic weighing of the errors. In the literature, forecast mean squared error has been used to refer to the mean of several MSPEs, which are based on bootstrapping.	- Sample level; - In-sample; - Observed expenses as reference; - Absolute measure;	**13. MSPE / MSE**	75
		- Sample level; - Out-of-sample; - Observed expenses as reference; - Absolute measure;	**14. MSPE / MSE**	51, 74
		-Subsample level - Out-of-sample - Observed expenses as reference; - Absolute measure;	**15. MSPE / MSE**	74
5. MSPE / MSE per decile of observed expenses	The MSPE has been calculated for deciles of (observed) expenses.	- Subsample level; - Out-of-sample; - Observed expenses as reference; - Absolute measure;	**16. MSPE / MSE per decile of observed expenses**	51
6. Root Mean Squared Error (RMSE) or RMSE to the mean	This measure is the root of the MSPE; i.e. the root of the mean of the squared deviations of predicted expenses from observed expenses. It is a quadratic score, which gives higher weight (penalty) to high deviations from observed expenses, followed by taking the root of them. The implicit weighing of prediction errors makes this measure sensitive to outliers.	- Sample; - In-sample; - Observed expenses as the reference; - Absolute measure;	**17. RMSE**	58
		Sample; - Out-of-sample; - Observed expenses as the reference; - Absolute measure;	**18. RMSE**	54
		- Sample; - In-sample; - Observed expenses as reference; - Relative measure;	**19. RMSE to the mean**	15 Remarks: RMSE compared to the average observed expenses, expressed in percentages (=standardized)

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
7. (conventional) PR / PTOR / CR (predictive ratio, or predictive-to-observed ratio, cost ratio)	A Predictive Ratio (PR), also referred to as Predictive-to-Observed Ratio, is the ratio of predicted expenses to observed expenses. The Cost Ratio (CR) is the reciprocal of the PR; i.e. the ratio of observed expenses to predicted expenses.	- Sample level; - In-sample; - Observed expenses as reference; - Relative measure;	**20. PR / PTOR / CR**	37, 39, 47, 55, 57, 63, 66
		- Sample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**21. PR / PTOR / CR**	19, 25, 28, 34, 51, 54, 56, 62, 65, 79, 80
		- Subsample level; - In-sample; - Observed expenses as reference; - Relative measure;	**22. PR / PTOR / CR**	4, 7, 11, 13, 18, 22, 37, 39, 41, 47, 55, 57, 58, 60, 63, 66, 73
		- Subsample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**23. PR / PTOR / CR**	1-3, 16, 18, 23, 25, 28-29, 31, 34, 35, 52, 54, 62, 65, 67, 79-81
		- Subsample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**24. Mean PR / PTOR / CR**	27, 35, 43 Remarks: Mean over PRs for multiple 'bootstrapped' groups or samples; Availability of a validation sample or external data is not necessary. Since bootstrapping creates multiple overlapping validation samples. The validation sample cannot be fully considered to be an independent sample, since they have observations in common with the in-sample.

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
8. PR per decile or quintile or other cost categories of observed expenses or predicted expenses	The PR calculated for groups representing a decile, a quintile, or another cost interval of observed or predicted expenses. Observed expenses or predicted expenses are first sorted in ascending order and based on this distribution groups are defined for which the PRs are calculated.	- Subsample level; - In-sample; - Observed expenses as reference; - Relative measure;	**25. PR per decile/quintile of observed expenses**	15, 57, 49, 50 Remarks: PRs for groups of observed expenses; Availability of (external) information to define groups is not necessary, since observed expenses are used to construct groups.
		- Subsample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**26. PR per decile/quintile of observed expenses**	9, 25, 51, 54 Remarks: PRs for groups of observed expenses; Availability of additional (external) information to define groups is not necessary, since observed expenses are used to construct groups. Note that an validation-sample needs to be available.
		- Subsample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**27. PR per decile/quintile/ cost category of predicted expenses**	47, 55, 66 Remarks: PRs for groups of predicted expenses; Availability of (external) information to define groups is not necessary, since predicted expenses are used to construct groups.

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
9. Predicted expenses to average observed expenses (E/A-ratio)	Calculated as the predicted expenses divided by average observed expenses.	- Sample level; - Out-sample; - (Average) observed expenses as reference; - Relative measure;	**28. E/A-ratio**	72 Remarks: The E/A-ratio at the sample level equals the PR at the sample level.
		- Subsample level; - In-sample; - (Average) observed expenses as reference; - Relative measure;	**29. E/A-ratio**	64
		- Subsample level; - Out-of-sample; - (Average) observed expenses as reference; - Relative measure;	**30. E/A-ratio**	65
10. Average Percentage Prediction Error Rate (APPER)	This measure is calculated as follows: the 'percentage prediction error' per group is calculated, which is the percentage that the 'predicted ratio' deviates from the 'observed ratio', or vice versa. Then the average of these 'percentage prediction error' is the average prediction error. The 'predicted ratio' is the predicted expenses divided by average predicted expenses and the 'observed ratio' is observed expenses divided by average observed expenses.	- Subsample level; - In-sample; - Observed expenses as reference; - Relative measure;	**31. APPER**	53, 72
		- Subsample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**32. APPER**	42 Remarks: In literature, also referred to as 'relative error'. The reference is the observed ratio or the predicted ratio in a sample or benchmark population, instead of individual observed expenses.

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
11. Mean Deviation Score (MDS)	The MDS is calculated as: first, the predictive ratio is calculated for several groups across the population. Then 1 minus the predicted ratio for all groups, followed by summing this difference across all groups, and dividing this by the number of groups and multiplying by 100; i.e. the average deviation of the PR from one for the all groups is calculated, expressed in percentages.	- Subsample; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**33. MDS**	3 Remarks: Ash et al. 2005 also calculated a weighted version of the MDS. This variant has not been treated as a separate measure, since the weights are explicitly chosen by the researcher and the measurement-specific properties are exactly the same as the conventional MDS.
12. (conventional) MPE / ME / MR / profit-loss / under-overpayment / forecast bias (mean prediction error / mean error / mean result)	The MPE is based on the difference between predicted expenses minus observed expenses. It can be computed on the full sample and subsample level. Given an OLS-model and expenses predicted on the in-sample, the MPE at the sample level is not meaningful, since the MPE-value would equal zero. The MPE at the sample level at an out-of-sample does not necessarily have to equal to zero. The MPE on the subsample level can be very useful and has been applied regularly.	- Sample level; - Out-of-sample; - Observed expenses as reference; - Absolute measure;	**34. MPE**	25
		- Subsample level; - In-sample; - Observed expenses as reference; - Absolute measure;	**35. MPE**	5, 11, 12, 36, 71

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
12. (conventional) MPE / ME / MR / profit-loss / under-overpayment / forecast bias (mean prediction error / mean error / mean result)		- Subsample level; - Out-of-sample; - Observed expenses as reference; - Absolute measure;	**36. MPE**	8, 9, 24, 25, 48
		- Sample level; - Out-of-sample; - (Sum of) observed expenses as reference; - Absolute measure;	**37. Gross profit**	68 Remarks: The total revenue of a plan (=group) is equal to the sum of premiums for all individuals in the study population. Total expenses are defined as the sum of the observed expenses over these individuals.
		- Subsample level; - In-sample; - Other reference than observed expenses; - Absolute measure;	**38. Normalized MPE**	70 Remarks: Normative expenses are the reference point, with normative expenses being predicted by another model; Availability of information to predict individual normative expenses.
		- Subsample level; - Out-of-sample; - Other reference than observed expenses; - Absolute measure;	**39. Mean predictable profit/loss**	44 Remarks: Predicted expenses are the reference point, based on a more advanced model; Availability of information to predict a more advanced model.

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
13. Average percentage prediction error (APPE)	Calculated as: the MPE divided by (mean) predicted or observed expenses of an insurer or other group ; i.e. it is the MPE as a percentage of predicted expenses or (mean) observed expenses.	- Sample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**40. APPE**	68 Remarks: In the literature, also referred to as 'gross profit rate' or 'percentage reduction in selection profit'.
		- Subsample; - In-sample; - (Average) observed expenses as reference; - Relative measure;	**41. APPE**	76 Remarks: In the literature, also referred to as the 'percentage difference'. This measure has been calculated as the MPE as a percentage of average predicted expenses, plus one.
		- Subsample; - Out-of-sample; - Observed expenses (average); - Relative measure;	**42. APPE**	8, 77 Remarks: In the literature, also referred to as 'average percentage under and overpayment'.
14. MPE per decile of predicted or observed expenses	This is the MPE for deciles of predicted or observed expenses.	- Subsample level; - In-sample; - Observed expenses as reference; - Absolute measure;	**43. MPE per decile of predicted expenses**	26 Remarks: Groups are based on deciles of predicted expenses.
		- Subsample level; - Out-of-sample; - Observed expenses as reference; - Absolute measure;	**44. MPE per decile of observed expenses**	25 Remarks: Groups are based on deciles of observed expenses.

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
15. Number and expenses for individuals with MPE below, within, or above X% of predicted expenses	The MPE is calculated. Then, this MPE is used to examine how many individuals have a MPE above a certain percentage of predicted expenses, together with the expenses of these individuals.	- Subsample level; - Out-of-sample; - Observed expenses as reference; - Absolute measure;	**45. Number and money amounts for individuals with MPE below, within, and above X% of predicted expenses**	48
16. MPE of one group divided by MPE for another group (= plan liability ratio)	This measure is calculated as: total observed expenses for each individual subtracted from risk-adjusted payments (i.e. predicted expenses) for each individual. What remained was the plan liability for each beneficiary; i.e. this is the profit or loss for each beneficiary. Then they calculated the ratio of mean adjusted plan liability of subsidy to non-subsidy beneficiaries for each risk decile. In the literature, this measure has also been referred to as 'plan liability ratio'.	- Subsample level; - In-sample; - Observed expenses as reference; - Absolute measure;	**46. MPE of a group divided by MPE of another group**	33 Remarks: Note that the MPE of a group is also used as a reference. Also referred to as 'plan liability ratio'.
		- Subsample level; - In-sample; - Observed expenses as reference; - Relative measure;	**47. MPE of one group divided by MPE for another group (=plan liability ratio)**	33 Remarks: Note that the MPE of a group is also used as a reference.
17. MPE for a group multiplied by the number of individuals within this group (='Total' financial gain/loss)	This measure is calculated as the average of the difference between mean observed expenses for a specific group and the mean predicted expenses for this group (= MPE), and then multiplying this average by the number of individuals within this group. This measure is also called the 'total financial gain or loss'.	- Subsample level; - Out-of-sample; - Observed expenses as reference; - Absolute measure;	**48. MPE for a group multiplied by the number of individuals within this group (='Total' financial gain/loss)**	52
18. Total profit	'Total profit' is calculated as the observed expenses minus the predicted expenses by the RE-model, summed for all individuals across all insurers (= groups).	- Sample level; - In-sample; - Observed expenses as reference; - Absolute measure;	**49. Total profit**	20

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
19. Percentage of individuals with MPE greater or smaller than a certain percentage level	This measure is calculated as the percentage of individuals whose observed total expenses were under-predicted or over-predicted (expressed in MPE) by more than a certain (chosen) percentage, which is used as a 'critical threshold value'.	- Subsample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**50. Percentage of individuals with MPE greater or smaller than a certain percentage level**	24
20. Range of average per-individual profit	'Range of average individual profit' is calculated as the average per-enrollee profit or loss (which is the MPE; i.e. predicted expenses minus observed expenses) across all health insurers under different models. The maximum and minimum value for each model is reported.	- Sample level; - Out-of-sample; - Observed expenses as reference; - Absolute measure;	**51. Range of average per-individual profit**	23
21. Relative MPE	This measure is calculated as: [the aggregate difference between revenues (=predicted expenses) and expenses (=observed expenses) by a model of a group to another group] minus [the aggregate difference between revenues and expenses by another model of a group to another group] Thus, two models are compared in terms of differences in revenues and expenses between the groups. In the literature, also referred to as the 'relative revenues'.	- Subsample level; - In-sample; - Observed expenses as reference. -Absolute measure;	**52. Relative MPE (= relative revenues)**	36

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
22. Percentage distribution of the error	The 'percentage distribution of the error' provides information about the distribution of the prediction error; i.e. how the deviations of predicted expenses to the reference are distributed among the population or groups. The distribution of the error can be divided into several intervals (for example percentiles or other groups), for each interval the percentage of individuals within this percentile can be presented. Furthermore, some descriptive statistics can be calculated for the distribution of the error; e.g. mean, standard error, minimum or maximum value, in order to indicate how the deviations from the reference are distributed among the population. But, these descriptive statistics are not included in our review.	- Subsample level; - In-sample; - Observed expenses as reference; - Relative measure;	**53. Percentage distribution of error**	38
		- Subsample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**54. Percentage distribution of error**	63
23. Forecast bias	In the literature, the term 'forecast bias' has been used to refer to the mean of bootstrapped 'MPE's' or forecast residuals. Forecast bias is calculated as: the average of the 500 mean forecast residuals, whereby the mean forecast residual is the difference between the average predicted expenses and average observed expenses. This measure has been calculated using bootstrapping techniques.	- Sample level; - Out of-sample; - (Average) observed expenses as reference; - Absolute measure;	**55. Forecast bias**	74, 14
		- Subsample level; - Out of-sample; - (Average) observed expenses as reference; - Absolute measure;	**56. Forecast bias**	74, 14

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
24. (conventional) MAPE (mean absolute prediction error / mean absolute result / mean absolute error)	The MAPE provides information about the absolute difference between predicted expenses and observed expenses. This measure assigns equal weight to each prediction error, making this measure less sensitive to large errors (outliers) than measures based on squared errors. In the literature, they also use the term mean absolute error' or 'mean absolute result'. Some studies computed the difference between observed and predicted expenses, but this is similar to predicted minus observed expenses when the absolute value of those differences have been taken. The MAPE is expressed in money amounts. Its value can be any positive number. A smaller value of the MAPE means that the model is better able to predict individual expenses than a model with a higher MAPE-value.	- Sample level; - In-sample; - Observed expenses as reference; - Absolute measure.	**57. MAPE / MAR / MAE**	4-6, 11, 18, 32, 40[d], 68, 71
		- Sample level; - Out-of-sample; - Observed expenses as reference; - Absolute measure;	**58. MAPE / MAR / MAE**	3, 9, 18, 23, 25, 56, 59, 62, 67, 68, 71
		- Subsample level; - Out-of-sample; - Observed expenses as reference; - Absolute measure;	**59. MAPE / MAR / MAE**	1, 24, 25, 62
		- Sample level; - In-sample; - Other reference than observed expenses; - Absolute measure;	**60. Normalized (sign-reversed) MAR**	70 Remarks: Normative expenses are used as the reference point; Availability of data to calculate normative expenses.
		- Sample level; - In-sample; - Other reference than observed expenses; - Absolute measure;	**61. MAPR**	5, 6, 44 Remarks: Other reference than observed expenses; predicted expenses as the reference (a more advanced model); Availability of data to estimate a more advanced model and predict expenses for each individual based on this model.

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
25. Mean Absolute Percentage Prediction Error (MAPPE)	The MAPPE is the MAPE-value as a percentage of mean observed or predicted expenses.	- Sample level; - In-sample; - Observed expenses as reference; - Relative measure;	**62. MAPPE**	15 Remarks: MAPE as a percentage of mean observed expenses; Relative measure, expressed in percentages.
		- Sample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**63. MAPPE**	16 Remarks: MAPE as a percentage of mean observed expenses; Relative measure, expressed in percentages.
		- Subsample level; - In-sample; - Observed expenses as reference; - Relative measure;	**64. MAPPE**	11 Remarks: MAPE as a percentage of mean observed expenses; Relative measure, expressed in percentages.
		- Subsample level; - Out-of-sample; - Observed expenses as reference; - Relative measure.	**65. MAPPE**	1 Remarks: MAPE as a percentage of mean predicted expenses; Relative measure, expressed in percentages.
26. MAPE per decile of observed expenses	This is the MAPE for deciles of observed expenses.	- Subsample level; - Out-of-sample; - Observed expenses as reference; - Absolute measure;	**66. MAPE / MAR / MAE per decile of observed expenses**	9, 25, 56 Remarks: MAPE per decile of observed expenses; Absolute measure, expressed in money amounts.

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
27. Reduction in MAPE	The reduction in MAPE for deciles of predicted expenses are based on the MAPE of a model compared to those of another model. This measure indicates the extent to which a model has been improved, compared to a reference model.	- Subsample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**67. Reduction in MAPE / MAR / MAE**	56 Remarks: The MAPE of a model for one group is calculated and the MAPE for the same group but based on another model. The MAPE for other groups are compared to this 'reference group'. Observed expenses are used as the reference point for calculating the prediction error; Absolute measure, expressed in money amounts; It is at least required to estimate two models and define groups.
28. 'Standardized' MAPE	To compute the 'standardized' MAPE, the observed and predicted expenses have first been standardized, which means that the observed and predicted expenses have been divided by the observed and predicted value of a reference group; i.e. the MAPE-value of one of the groups is set equal to 1.0. Then, the absolute difference between the standardized observed and predicted expenses for each other group is computed, followed by taking the average of these absolute differences. This measure is standardized in order to indicate the dispersion between different groups.	- Subsample level; - Out-of-sample; - Observed expenses as reference; - Relative measure;	**68. Standardized MAPE / MAR / MAE**	30 Remarks: The MAPE of a group is used as the reference. The MAPE of other groups are compared to this reference. Note that observed expenses are used as the reference to compare predicted expenses with.
29. Mean Absolute Deviation (MAD)	The MAD is calculated as: the mean of the absolute difference between predicted expenses and average observed expenses.	- Sample; - Out-of-sample; - (Average) observed expenses as reference; - Absolute measure;	**69. MAD**	40 Remarks: In literature, also referred to as 'mean financial transfer'.

Table A.2.1: (continued)

CLUSTERED MEASURES **Table 2.2 in the main text.** Measures after clustering based on similarities and differences in properties of the measures and four distinguished variations in analytic method (which are presented in the next column).		**Variations in analytic method**	**TREATING EACH VARIANT A MEASURE SEPARATELY** Identified measures before clustering, taking into account differences in the way measures have been applied, according to the four distinguished variations (presented in the previous column).	
Measure	**Short description**		**Measure**	**Study** [a,b]
30. Cumming's Prediction Measure	The CPM is a goodness-of-fit measure. Like the R^2, the CPM is expressed on a standardized scale from 0 to 1. However, the CPM is measured in absolute prediction errors rather than (individual) prediction errors. Due to this property, the CPM-value is less sensitive to large prediction errors, such as outliers.	- Sample level; - In-sample; - Observed expenses as reference; - Relative measure.	**70. (Estimated) CPM**	18, 15[e], 11
		- Sample level; - Out-of-sample; - Observed expenses as reference; - Relative measure.	**71. (Validated) CPM**	3, 9, 18

Footnotes Table A.2.1:

a. See reference list in Appendix 2.2 for the number that corresponds to each study.
b. For example, "1-6" means study 1, 2, 3, 4, 5, *and* 6 in the reference list.
c. These studies have used the maximum R^2 estimated in another study. They mention that the autocorrelations– and thus, the max R^2 – are roughly the same as those in the other study; however, strictly spoken, another dataset have been used.
d. van Kleef & van Vliet (2010) have also calculated the marginal increase in MAD and MAR, which both have not been treated as a separate measure in this table, since this is the MAD or MAR of one model minus the MAD or MAR of another model.
e. Chalupka (2010) did not explicitly refer to the CPM, but they calculated the 'sum of absolute errors' of an RE-model to no RE-model (= average expenses), making this measure equivalent to the CPM.

Chapter 3

Estimating the Potential Selection Profits

ABSTRACT

Risk equalization (RE) models attempt to make risk selection unprofitable by paying according to the expected expenses for selective groups of interest. This chapter develops and empirically applies three methods to estimate the potential selection profits on multiple groups simultaneously. These methods create mutually exclusive groups from a pre-defined set of overlapping groups. Although the three methods yield different estimates of the potential selection profits in absolute money amounts for each RE-model, they come to the same conclusion about which RE-model yields the largest reduction in the potential selection profits. The methods that are developed in this study are generally applicable for evaluating any RE-model, conditional on any set of pre-defined groups of interest. Aggregated residual expenses on overlapping groups, as used in previous studies, over-estimate the potential selection profits under a given RE-model. However, usage of overlapping groups instead mutually exclusive groups does not lead to another conclusion about which RE-models yields the largest reduction in the potential selection profits.

§ 3.1 INTRODUCTION

Several countries world-wide, including Belgium, Germany, Israel, the Netherlands, Switzerland, and the U.S., use a risk equalization (RE) system in order to compensate health insurers for predictable variation in individuals' healthcare expenses (van de Ven & Ellis, 2000; Kautter et al., 2014). A RE system provides risk-adjusted payments to insurers that are calculated by a prediction model that uses several risk factors to predict individuals' healthcare expenses. 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 to achieve a level playing field for health insurers. Risk selection is a potential threat to solidarity, efficiency, and quality of care (Baumgartner & Busato, 2012; Beck, 2003; Frank et al., 1998; van de Ven et al., 2007; von Wyl & Beck, 2015; see also § 1.2.2). For example, insurers can engage in risk selection by not contracting high-quality care or providing poor services to selective high-cost groups (Newhouse, 1996; van de Ven et al., 2007).

An appropriate method to investigate the extent to which RE-models mitigate financial incentives for risk selection is to assess the predictive performance of these models for *selective* groups[1] (van Veen et al., 2015a). These selective groups are defined by risk factors that can be derived from any type of information that may be used by the insurer for exploiting risk selection[2]. For example, an insurer could use the same type of information as used by the regulator for estimating the RE-model to define selective groups that are not explicitly used in the RE-model; for example, some interaction terms between risk adjusters, or an insurer could use external information that is not used for estimating the RE-model, such as health survey information. Under this asymmetric information structure, the insurer can exploit the extra information to obtain its own expectation about the profitability of specific groups of interest under a given RE-model. Large average residual expenses (= observed expenses minus RE-predicted expenses) for selective groups, especially when they persist

[1] Van Veen et al. (2015a) advocate that analyses on the full sample, random groups, or random insurers' portfolios level are not adequate for measuring financial incentives for risk selection because the outcomes of these analyses can be the net effect of aggregation of significant under- and over-compensations for selective groups.

[2] It is worth noting that groups that are identical or (very) closely related to the risk adjusters in the RE-model that is evaluated are not of interest when this model is estimated by Ordinary Least Squares (OLS), which is so far the conventional estimation technique (van Veen et al., 2015a). This is because any RE-model that is estimated by OLS adequately predicts average expenses for the groups that are explicitly included in the model. This is not the case when another estimation technique than OLS is used, whereby average predicted expenses does not have to equal average observed expenses per group that is explicitly included in the model. In that case, selective groups of interest can be the same as the risk adjusters included in the model. Consequently, an insurer does not have to exploit extra information in order to engage in risk selection but using the same type of information as used by the regulator for defining the risk adjusters in the RE-model may be sufficient to find selective groups of interest with large residual expenses under a given RE-model.

over time, create financial incentives for risk selection. An insurer may realize substantial profits when actions by the insurer or consumer result into disproportionate enrollment of individuals for whom the RE-model over-predicts expenses; i.e. the favorable group, and/or *dis*enrollment of individuals for whom the RE-model under-predicts expenses; i.e. the unfavorable group[3]. If the RE-model predicts expenses adequately for any selective group of interest, then risk selection is not profitable.

Over the past decades, several studies have compared the predictive performance of alternative RE-models on several selective groups of interest by analyzing average residual expenses per model for each group (e.g. Ash & Byrne-Logan, 1998; Ash et al., 2005; Pope et al., 2000a; van Kleef et al., 2012a, 2012b, 2013b; van Veen et al., 2015b). Since RE-models can perform differently on different groups, it is possible to obtain conflicting outcomes; i.e. a model may reduce average residual expenses for one group but may increase it for another (e.g. Fishman et al., 2003; Mark et al., 2003). For decision-making in such situations, it is very helpful if the potential selection profits for multiple groups simultaneously can be estimated for each RE-model that is evaluated.

To obtain an estimate of the potential selection profits for multiple groups simultaneously it is important to create *mutually exclusive* groups (Ash et al., 2005; van Kleef et al., 2012a). Overlap between groups may especially occur for groups based on the presence of chronic conditions or healthcare utilization; for example, there may be a large overlap between groups that are defined by questions in a health survey about general health status, the presence of one or more chronic diseases, and the usage of certain healthcare facilities. Aggregating residual expenses on overlapping groups yields a biased estimate of the potential selection profits, because individuals who are in multiple groups have a relatively large impact on the estimate; e.g. individuals with multiple chronic conditions generally have high above-average expenses and large residual expenses (see § 3.3 for an empirical illustration).

Mutually exclusive groups could be created by defining groups based on each possible combination of the risk factors. For example, usage of two dichotomous risk factors results into four distinctive groups. However, when many risk factors are used this method is no longer practical and some groups may no longer be meaningful because they are very small, whereby random error can play a large role. For example, 45 dichotomous risk factors that are used in some Dutch studies theoretically result into 2^{45} distinctive groups per RE-model (van Kleef et al., 2012a, 2012b; van Veen et al., 2015b).

[3] Van Kleef and colleagues customized Newhouse's definition for risk selection for a market with RE as 'actions other than risk rating per product by consumers and insurers with the intention and/or the effect that solidarity [i.e. the intended pooling of low and high risks] is not fully achieved' (Newhouse, 1996; van Kleef et al., 2013a).

The goal of this study is to develop and empirically apply three methods for creating mutually exclusive groups from a pre-defined set of overlapping groups in order to estimate the potential selection profits for these groups simultaneously under various RE-models (see § 3.2 for a detailed description). These methods make from multiple overlapping groups two distinctive groups in the population: i.e. the favorable group and unfavorable group. Analyzing average residual expenses of a given RE-model for these distinctive groups provides an estimate of the potential selection profits under this RE-model. We aim to provide insight into whether the pattern in the percentage reduction in potential selection profits under various RE-models is similar across different methods, which is of importance when alternative RE-models are evaluated in order to decide which model should be used in practice.

Our estimates reflect the *potential* selection profits that can be exploited by an insurer from a set of pre-defined overlapping groups of interest by rejecting all individuals in the unfavorable group or attracting all individuals in the complementary favorable group, under a given RE-model. We ignore transaction costs for engaging in risk selection and assume that the insurer is able to exploit the extra information optimally. Our estimates are conditional on the set of information that is available to create pre-defined overlapping groups and the definition of these groups.

It is worth noting that this study exclusively focusses on measuring potential selection profits under various RE-models for the purpose of evaluating these RE-models; we do *not* attempt to investigate to what extent insurers actually do engage in risk selection and how effective they are. Over the past decades, it has proven to be very difficult to show whether and how insurers exploit risk selection because of several methodological challenges. A few studies attempted to answer this question and concluded that risk selection is a real phenomenon (Yu et al., 2001; Newhouse et al., 2012, 2015; van Kleef et al., 2013a; von Wyl & Beck, 2015).

The assessment of the potential selection profits is of great policy relevance because it enables policymakers to monitor to what extent a given RE-model mitigates financial incentives for risk selection. This assessment may help them to decide which RE-model leads to the largest reduction in potential selection profits for a set of pre-defined groups of interest and so, which model should be used in practice. In principle, the methods applied here can be used for evaluating any RE-model and are generally applicable for any type of information that can be used to define selective evaluation-groups of interest.

The methods that are developed here are applied by using real individual-level administrative data from almost the entire Dutch population of insured and health survey data from a representative sample of this population. We preface this empirical analysis with describing the methods and outlining how we integrate prior knowledge about evaluating RE-models in order to estimate the potential selection profits for multiple groups simultaneously. Section 3.3 describes the data and method of the empirical analysis. Section 3.4

reports the empirical findings. Section 3.5 concludes and Section 3.6 discusses the methods and findings.

§ 3.2 HOW TO ESTIMATE THE POTENTIAL SELECTION PROFITS?

This study discusses and empirically tests three methods that can be used for creating mutually exclusive groups from a set of pre-defined overlapping groups, with the purpose to estimate the potential selection profits under various RE-models. In the literature, there may also be alternative methods that may be adequate for measuring potential selection profits[4]. However, application of the methods discussed here start with a set of pre-defined *overlapping selective groups of interest*, whereby the same set of information can be used. This way, they are extensions of the aforementioned commonly-used evaluation method of analyzing average residual expenses on separate groups under a given RE-model. The methods described here may be relevant in practice when various RE-models are evaluated on many different groups.

In the present study, we assume that insurers have extra information on the individual level that is exploited to define a set of selective groups. This is an extreme form of asymmetric information; in Section 3.6 we will discuss whether it is necessary that an insurer has this information in practice. Methodologically, it does not matter whether the risk factors that are used for defining the selective groups are derived from claims information in insurer's administrative files or information in health surveys, because in principle with any type of information selective groups of interest can be created. If the extra information to define the groups is only available for a representative sample of the population it is important to incorporate statistical uncertainties around average residual expenses for the mutually exclusive groups. The methods developed here are only relevant if there is some overlap between the pre-defined groups. In the extreme situation of no overlap, residual expenses for multiple groups can be aggregated immediately in order to obtain an estimate of the potential selection profits.

§ 3.2.1 Creating mutually exclusive groups by using a stepwise removal algorithm

A relatively simple method for creating mutually exclusive groups is imposing a hierarchy for assigning individuals to one of the pre-defined groups. This stepwise-removal (SR) algorithm has many similarities to another hierarchical classification algorithm that has

[4] The starting point in this study is that the extra information is exploited for defining a set of overlapping groups. This starting point is distinctive from Newhouse and his colleagues' method for estimating the potential selection profits (Newhouse et al., 1989). They have used the maximum R^2 as a starting point to examine the additional percentage of variance in observed expenses that is explainable by the insurer, with the aim to estimate the selection profits in money amounts.

been used in field of RE: the Diagnostic Cost Group Model (Ash et al., 1989; Ellis & Ash, 1995). The SR algorithm relies on the principle that the information on selective groups is used efficiently by first identifying the most (un)favorable group from a set of pre-defined groups, then the second most (un)favorable group and so on. Identifying the (un)favorable group can be done in different ways. Below we will discuss two criteria for determining the hierarchy of selecting the most (un)favorable groups from a set of pre-defined groups. Using a different criterion may lead to selecting other groups. In our empirical application we will examine to what extent another criterion results into a different estimate of the potential selection profits under the same RE-model.

The hierarchy is determined by an iterative process, consisting of three steps: (1), for each group, the value of criterion 'X' is calculated and all groups are sorted in ascending or descending order of this criterion, depending on whether the focus is on groups for which the RE-model under- or over-predicts expenses; (2), the group with the highest rank is selected and individuals in this group are removed from the sample; (3), the process starts at step one again for the remaining groups on the reduced sample, until pre-specified stopping rules are met. All individuals selected by the algorithm are classified to one overarching group; all remaining individuals in the sample form the complementary group. In this way, the SR algorithm makes from multiple overlapping groups two distinctive groups (i.e. the aggregated favorable and unfavorable group), which are then used for estimating the potential selection profits of an RE-model. Note that the use of the SR algorithm makes the interpretation of residual expenses for separate groups difficult, because residual expenses for groups further down in the hierarchy are conditional on not being assigned to previously selected groups.

A first criterion we apply here for determining the hierarchy of the SR algorithm is average residual expenses (*method 1*). This criterion may select (small) groups with high average residual expenses, which may have a relatively large impact on the estimate. If the focus is on identifying groups for which the RE-model under-predicts expenses, then the ranking of groups is based on the magnitude of average residual expenses, given that its value is positive and statistically significantly deviates from zero. Then, the algorithm stops when average residual expenses for the remaining groups do not statistically significantly deviate from zero or average residual expenses are negative because then the RE-model over-predicts expenses for these remaining groups. Inversely, if the focus is on identifying groups for which the RE-model over-predicts expenses, the ranking is based on groups with negative average residual expenses.

A second relevant criterion for determining the hierarchy is total residual expenses (*method 2*), which is defined as average residual expenses multiplied by the size of the group. This criterion is similar to the previous one, except that it incorporates the size of the group. Group size may play a role in addition to the magnitude of average residual expenses when the aim is to maximize total profit. Inherently, this criterion may overlook small groups with

(very) large average residual expenses. Further, this criterion may select fewer groups than the first criterion, because the chance of overlap is larger between the selected large groups and the remaining (smaller) groups. If the focus is on identifying groups for which the RE-model under-predicts expenses, the algorithm selects groups with statistically significantly positive total residual expenses and stops when total residual expenses for the remaining groups are negative or average residual expenses do not statistically significantly deviate from zero. Inversely, if the focus is on identifying groups for which the RE-model over-predicts expenses, the algorithm selects groups with negative total residual expenses and stops when the remaining groups have positive total residual expenses or average residual expenses do not statistically significantly deviate from zero.

Assessing model's predictive performance by using the aforementioned two criteria may lead to different estimates under the same RE-model, because each criterion may select different groups. On theoretical grounds, there are no clear arguments in favor of one of the criteria: (i), criterion 1 may select small (homogeneous) groups with high average residual expenses but ignores the size of this group; (ii), criterion 2 incorporates the size of this group but there is a chance of selecting large (heterogeneous) groups. Further, it is unclear how these criteria work out in comparison to a third method, which will be discussed below. Our empirical application will indicate to what extent using another method leads to a different estimate of the potential selection profit under the same RE-model.

When comparing the predictive performance of RE-models, the algorithm should be conducted per RE-model that is evaluated because residual expenses per RE-model may differ. The algorithm selects those groups for which an RE-model does not predict expenses adequately. Consequently, different groups may be selected when different RE-models are evaluated, given the same criterion for determining the hierarchy. However, this is not of particular interest when the purpose is to determine which RE-model yields the lowest potential selection profits, conditional on using the set of pre-defined groups optimally under a given RE-model and using the same criterion for determining the hierarchy.

§ 3.2.2 Creating mutually exclusive groups by using stepwise regression analysis

A completely different method than the aforementioned two methods is to use stepwise regression analysis techniques for creating mutually exclusive groups (*method 3*). With this method, the pre-defined groups are used as explanatory variables in order to predict residual expenses for each individual under a given RE-model that can be expected when exploiting the extra information. Model's predictions are used to classify individuals in the sample into the favorable or unfavorable group. Average residual expenses of the RE-model (and *not* predicted residual expenses) for these distinctive groups provide an estimate of the potential selection profits. This method has some similarities to studies that have used a selection model with a more advanced set of risk adjusters than those included in the RE-model for determining the group of individuals for whom RE-predicted expenses exceed

insurer-predicted expenses and the group of individuals for whom RE-predicted expenses fall behind insurer-predicted expenses (e.g. Lamers, 2001; Shen & Ellis, 2002; Donato & Richardson, 2006; Eggleston & Bir, 2009). Just as those studies, we exploit extra information in order to obtain predictions of individuals' profitability, given the RE-model that is used. However, a crucial difference is that we apply stepwise regression analysis instead of conventional OLS analysis techniques with all variables included in the model. In this way, we use the extra information optimally by using a subset of all variables that best explains variability in residual expenses. All selected variables have statistically significant explanatory power and so, we reduce variance in model's predictions that may be caused by inclusion of irrelevant (i.e. not statistically significant) variables in the model.

Our method works as follows. First, residual expenses of a given RE-model are regressed on a set of dummy variables for pre-defined groups[5,6], using Ordinary Least Squares (OLS) with a stepwise variable selection method. OLS is used because this is the conventional estimation method for RE-models; though, other statistical model specifications could also be used that incorporate specific distributional properties of residual expenses, such as a high skewness (e.g. Duan et al., 1983; Manning & Mullahy, 2001; Manning et al., 2005). A stepwise variable selection method is used to select only those groups from a set of pre-defined groups that statistically significantly explain variation in residual expenses of an RE-model. For simplicity, we only incorporate main effects in the regression model and do not incorporate interaction effects between pre-defined groups. In this way, we use the same set of groups as the starting point for all three methods that are applied here. In practice, however, any interaction term that is of interest could be incorporated in the model. Second, the estimated coefficients are used to predict residual expenses for each individual in the sample. All individuals with positive predicted residual expenses are assigned to the unfavorable group; conversely, all individuals with negative predicted residual expenses are assigned to the (complementary) favorable group. These two distinctive groups are used for estimating the potential selection profits under a given RE-model, which will be done by aggregating residual expenses of the RE-model for each group and *not* predicted residual expenses of the regression model. Consequently, it is possible that individuals with negative residual expenses of an RE-model can be assigned to the unfavorable group because they have positive predicted residual expenses; and individuals with positive residual expenses

[5] Note that if you use exactly the same set of risk factors as used as risk adjusters in the RE-model when this model is estimated by OLS, then the stepwise regression model cannot find any statistically significant risk factor, resulting into zero potential selection profits.

[6] Here we assume that the extra information is used to define groups, even if this information is available on a continuous scale of measurement, such as individual-level prior years' expenses. Note that with this method (i.e. regression analysis) extra information that is measured on a continuous scale of measurement could be used. However, the other two methods in this study require that the extra information is measured on a categorical scale of measurement. Consequently, some information will be lost when information on a continuous scale of measurement needs to be aggregated to categories.

can be assigned to the favorable group because they have negative predicted residual expenses. The extent that this happens depends on the predictive power of the extra information that is exploited.

To illustrate this method, let us assume that we assess the predictive performance of an RE-model (Model A). For this model, predicted residual expenses are calculated by estimating the regression model. All individuals with positive predicted residual expenses are assigned to the unfavorable group because for those individuals an insurer expects that the RE-model under-predicts expenses, given that Model A is used. Inversely, all individuals with negative predicted residual expenses form the favorable group. If the extra information is positively correlated and statistically significantly explains variability in residual expenses of the RE-model, the insurer obtains a higher estimate of average residual expenses for the unfavorable group than average residual expenses of the RE-model for this group. If the extra information has less explanatory power because the RE-model uses a more advanced set of risk adjusters (i.e. RE-residual expenses reduces), an insurer obtains a lower estimate of average predicted residual expenses for this group than exploiting information with high explanatory power. If the extra information has no explanatory power, which happens when the RE-model perfectly predicts expenses for the pre-defined groups, predicted residual expenses are zero and so, there are no potential selection profits. Note that the total sum of predicted residual expenses for the favorable and unfavorable group equals zero because the RE-model and the regression model for predicting residual expenses are both estimated by OLS.

Let us now assume that Model A is improved by including a new risk adjuster (Model B). In this situation, it is of interest to compare Model A to Model B in order to examine to what extent the potential selection profits change as a result of including a risk adjuster. Just as for Model A, predicted residual expenses are calculated under Model B. If predicted residual expenses of Model A and B are graphically depicted in a figure with individuals' predicted residual expenses sorted in descending order on the X-axis per RE-model than the curve of Model B is less steep than for Model A. It is expected that the curve flattens as the predictive performance of the RE-model increases (i.e. RE-residual expenses reduces), because the risk factors of the pre-defined groups can explain less variance in residual expenses of the RE-model, given a positive correlation between the risk factors and residual expenses. Note that a horizontal curve equaling the X-axis implies that the information on selective groups cannot explain any variance in residual expenses of the RE-model and so, predicted residual expenses are zero. In this case the RE-model perfectly predicts expenses for all pre-defined groups and so, there are no potential selection profits. In this example, Model B is preferred above Model A based on its predictive performance, because the risk factors can explain less variance in RE-residual expenses and so, Model B yields lower potential selection profits, given the same set of information for defining the selective groups. In our empirical application we will graphically depict predicted residual expenses of the RE-models that are

evaluated in this study in order to demonstrate to what extent the potential selection profits that are expected by the insurer – note that this is not the same as our estimate that uses RE-residual expenses – change as a more advanced set of risk adjusters is used.

§ 3.2.3 Estimate of the potential selection profits

Based on the mutually exclusive groups as identified by each of the three aforementioned methods, the potential selection profits under a given RE-model can be calculated as the sum of residual expenses of this RE-model over all individuals that are assigned to the favorable or unfavorable group, respectively, divided by the total number of insured-years[7] in the sample. The magnitude of this estimate indicates the average *potential* selection profits under a given RE-model for one year per individual that can be obtained by attracting the total favorable group or by rejecting the total unfavorable group before the next contract year, conditional on the set of pre-defined groups. The magnitude of the potential selection profit for the unfavorable group and the complementary favorable group is equal to each other because the combined average residual expenses equals zero (i.e. a property of OLS). The closer the profits are to zero, the higher models' predictive performance for the pre-defined groups, whereby zero profits implies perfect model fit[8].

§ 3.2.4 Which method to be used?

An important difference between the stepwise removal algorithm and the stepwise regression method is how they create mutually exclusive groups. The stepwise removal algorithm sequentially removes overlap between groups, making it possible to determine which groups are still statistically relevant after removing overlap between pre-defined groups. The stepwise regression method, however, predicts individuals' residual expenses based on a set of groups that best explains variance in residual expenses of a given RE-model. Here the pre-defined groups of interest are used all together at the same time, whereby the focus is on prediction and not on hypothesis testing or causal interpretation. Consequently, it is not possible to determine which groups are statistically relevant when overlap between groups is removed, because the total set of estimated coefficients is of interest, regardless of the sign of the coefficients. To conclude, the stepwise removal algorithm and stepwise regression method completely differ in how they exploit the information on groups. There are no clear theoretical arguments in favor of one of the methods.

---

[7] In this study, not all individuals are enrolled the full contract period of one year and therefore, we use the total number of insured-years as the divisor because residual expenses are annualized in the calculations. Note that if all individuals in the sample are enrolled the full contract period, the number of individuals is equal to the number of insured-years.

[8] Note that zero profits would imply that the RE-model perfectly predict expenses for all pre-defined selective groups. In this case, the methods cannot select any group and stop immediately.

In the present study, we are interested in how these methods work out on real data. In particular, we empirically test which method yields the largest potential selection profits under a given RE-model and whether different methods lead to different conclusions about the relative predictive performance of an RE-model when alternative RE-models are evaluated, conditional on the same set of pre-defined groups. The method leading to the largest profit for the same RE-model may be preferred because then the extra information is used optimally under this RE-model. Further, if the percentage reduction in potential selection profits under a given RE-model is similar across different methods, it does not matter which method is used when the purpose is to determine which model is the preferred one to be used in practice based on its predictive performance; and *not* to conclude how large the potential selection profits are in absolute money amounts.

§ 3.3 EMPIRICAL ANALYSIS

§ 3.3.1 Administrative data and health survey data

Administrative data from almost the entire Dutch population of insured in 2011 (N = ~16.7 million) was used to estimate several RE-models. This dataset contained information on individual total observed healthcare expenses, demographics, and all other risk adjusters in the Dutch RE-model, including age interacted with gender (40 risk classes), source of income interacted with age (18 risk classes), region (10 risk classes), socioeconomic status interacted with age (12 risk classes), pharmaceutical cost groups (PCGs, 24 risk classes), diagnostic cost groups (DCGs, 16 risk classes), multiple-year high cost groups (MHC-groups, 7 risk classes), and durable medical equipment groups (DME-groups, 5 risk classes). A more detailed description of these risk adjusters is well-documented elsewhere: see Table A.1, page 255 (Eijkenaar et al., 2013). Total healthcare expenses were the expenses included in the basic benefit package, except mental healthcare services[9]. Total expenses were annualized and weighted by the fraction of the year an individual was enrolled: e.g. an individual who was enrolled 6 months and had healthcare costs of € 500 was given a weight of 0.5 and € 1000 annualized total healthcare expenses.

A Dutch health survey from 2010, "Gecon", was used to create selective groups. This survey is conducted each year by "Statistics Netherlands" on a representative sample of the Dutch population, aiming to collect information on self-reported health and healthcare utilization. The study population consists of children and adults in private households. So far, survey information is not used for defining risk adjusters in any RE-model – except for the Israeli RE-model (Shmueli, 2015) – and it is questionable whether this will happen

[9] Mental healthcare expenses were excluded because in the Netherlands a separate RE-model with different risk adjusters is used for these expenses.

because of several practical problems; e.g. surveys can be costly, response rates can be unacceptably low, and self-reported measures may be vulnerable to manipulation (Hornbook & Goodman, 1996; Stam et al., 2010b; Yu & Dick, 2010). As long as survey information is not explicitly used as a risk adjuster in RE-models, it is very useful for evaluating RE-models. It is also possible to monitor the potential selection profits under various RE-models over time by using the same definition for the set of pre-defined groups, given that the same survey is conducted over time.

§ 3.3.2 Estimation of the risk equalization models

To apply the three methods, we estimated eight RE-models. We started with a model including only an intercept to indicate the situation of no RE (Model 0). Then, we sequentially added the following risk adjusters to the model: age interacted with gender (Model 1), region (Model 2), source of income interacted with age (Model 3), PCGs (Model 4), DCGs (Model 5), socioeconomic status interacted with age (Model 6), MHC-groups (Model 7), and DME-groups (Model 8). These models approximately represent the models that have been successively used in the Netherlands in the period 1993 to 2014 (van Kleef et al., 2012a).

All RE-models regressed annualized individual total healthcare expenses on dummy variables for the risk adjusters in the model, using OLS with a weight for the enrollment period. To prevent overfitting, all RE-models were estimated on a random half of the total administrative dataset; i.e. the estimation sample, and the estimated coefficients were used to predict expenses in the remainder half of this dataset; i.e. the validation sample. All individuals that were respondents to the survey were first assigned to the validation sample in order to make maximum use of the survey data. All remaining individuals in the administrative dataset were randomly assigned to one of the samples. The validation sample was merged with the survey results at the individual level by an anonymous identification key. This split sampling procedure did not yield bias in the representativeness of the administrative samples (Table 3.1)[10].

On the validation sample, the R-squared of the RE-models 0 to 8 was 0%, 4.4%, 4.4%, 5.1%, 11.3%, 20.7%, 20.8%, 24.2%, and 24.3%, respectively. Cumming's Prediction Measure and the Mean Absolute Prediction Error showed a similar pattern in models' predictive per-

[10] More descriptive statistics of the total administrative dataset and all samples are presented in Appendix 3.1.

formance (Appendix 3.2)[11,12]. All these measures-of-fit indicate that the inclusion of more risk adjusters in the RE-model has substantially increased models' predictive performance on the full sample.

§ 3.3.3 Definition of the set of pre-defined selective groups

Questions about self-reported health and healthcare utilization were used to create 46 selective groups, using similar definitions as those used in prior research (Stam et al., 2010b; van Kleef et al., 2013b; van Veen et al., 2015b). Examples were "How do you rate your health status?" and "Do you have one of the following diseases or chronic disorders?". Table A.2 describes all pre-defined groups. These groups consisted of an over-representation of individuals with above-average expenses. Average expenses for each group were based on the total annualized expenses of each individual within this group for treating all diseases of this individual; e.g. for the group of individuals with Diabetes it was not only the expenses of treating this specific disease. The RE-models that are evaluated on average under-predict expenses for the pre-defined groups.

§ 3.3.4 Representativeness of the survey sample

To obtain a representative sample of the population, "Statistics Netherlands" followed several procedures. One of them is the usage of a mixed-method design in order to avoid selective non-response[13]. Further, they conducted the survey in different periods of the year in order to prevent bias resulting from seasonal trends; e.g. a higher prevalence of depression in autumn and winter time[14]. Table 3.1 shows that the survey sample (N = 16,141) is considered reasonably representative for the Dutch population in terms of the prevalence of a PCG, DCG, MHC-group, and DME-group, except for those individuals in nursing homes or other institutions because these individuals are excluded from sample selection (Table 3.1). Consequently, average age in this sample is lower than average age in the population.

---

[11] See van Veen et al. (2015a) for a description of these measures-of-fit (Chapter 2).

[12] A plot with the relationship between the R^2 and the CPM and our estimates of the potential selection profits is provided in Appendix 3.3. This figure shows that the R^2 is a non-linear function of the potential selection profits: the first increases in percentage points in R^2 lead to a relatively large reduction in the potential selection profits and for each additional percentage point in R^2 the marginal reduction in the potential selection profits decreases. The CPM is a more linear function of our estimates of the potential selection profits.

[13] This mixed-method design includes the usage of "Computer Assisted Web Interviewing", "Computer Assisted Telephone Interviewing", and "Computer Assisted Personal Interviewing" for all questions and "Paper and Pencil Interviewing" and "Paper and Web Interviewing" for specific questions in an additional questionnaire for individuals older than 12 or 54 years.

[14] See website of "Statistics Netherlands" for a detailed description of all procedures that are followed for sample selection: www.cbs.nl.

Table 3.1: Descriptive statistics of the total administrative data from the Dutch population of insured in 2011, the estimation- and validation-sample of this administrative dataset, and the survey sample from 2010 [a]

	Administrative dataset			Health survey data	
	Total dataset [b]	**Estimation-sample** [b]	**Validation-sample**	**Merged survey sample** [c]	**Survey sample (after correction)** [c]
N(individuals)	16,688,961	8,327,580	8,361,381	16,141	15,604
N(insured-years)	16,438,958	8,201,696	8,237,262	16,067	15,535
Expenses					
Mean total observed healthcare expenses, in €'s (std.) [d]	1,785 (5,978)	1,783 (5,944)	1,786 (6,011)	1,766 (5,364)	1,720 (5,325)
Median total observed healthcare expenses, in €'s	445	445	446	444	428
Mean predicted expenses *Model 0*, in €'s [e]			1,786 (0)	1,786 (0)	1,786 (0)
Mean predicted expenses *Model 1*, in €'s [e]	n.a.	n.a.	1,785 (1,258)	1,752 (1,230)	1,708 (1,192)
Mean predicted expenses *Model 2*, in €'s [e]	n.a.	n.a.	1,785 (1,264)	1,746 (1,237)	1,701 (1,199)
Mean predicted expenses *Model 3*, in €'s [e]	n.a.	n.a.	1,784 (1,351)	1,721 (1,315)	1,673 (1,278)
Mean predicted expenses *Model 4*, in €'s [e]	n.a.	n.a.	1,785 (2,023)	1,724 (1,968)	1,672 (1,918)
Mean predicted expenses *Model 5*, in €'s [e]	n.a.	n.a.	1,786 (2,755)	1,749 (3,056)	1,698 (3,047)
Mean predicted expenses *Model 6*, in €'s [e]	n.a.	n.a.	1,786 (2,756)	1,744 (3,053)	1,694 (3,045)
Mean predicted expenses *Model 7*, in €'s [e]	n.a.	n.a.	1,787 (2,971)	1,762 (3,291)	1,712 (3,277)
Mean predicted expenses *Model 8*, in €'s [e]	n.a.	n.a.	1,786 (2,975)	1,764 (3,297)	1,713 (3,283)
Risk characteristics					
Mean age in years (std.)	40.1 (22.924)	40.1 (22.934)	40.1 (22.914)	39.7 (22.986)	39.0 (22.720)
Proportion male	0.493	0.493	0.493	0.486	0.490
Proportion classified in a PCG [f]	0.173	0.173	0.173	0.172	0.165
Proportion classified in multiple PCGs	0.035	0.035	0.035	0.035	0.032
Proportion classified in a DCG [g]	0.087	0.086	0.087	0.087	0.083
Proportion classified in a MHC [h]	0.058	0.058	0.058	0.059	0.056
Proportion classified in a DME [i]	0.008	0.008	0.008	0.009	0.009
Proportion classified in a PCG, DCG, MHC, and/or DME	0.220	0.220	0.221	0.220	0.212

Footnotes Table 3.1:

a. Statistics in this table are weighted for the enrolment period. To calculate the standard deviation (in parentheses), the sum of the weights minus one is used as the variance divisor.
b. RE-predicted expenses are not available for the total administrative dataset and the estimation-sample of this administrative dataset, because the coefficients of the RE-models were estimated on the estimation-sample and expenses were predicted on the validation-sample. The models were estimated by Ordinary Least Squares (OLS). Average predicted expenses by some models (slightly) deviate from average observed expenses in the validation sample, because of the usage of a split sampling approach. Further, average predicted expenses in the survey sample deviate from average predicted expenses in the population. To correct for these differences, residual expenses in the survey sample were calibrated in such a way that average residual for each model on this sample equals zero (i.e. individuals' residual expenses per RE-model were raised by a factor equaling average RE-predicted expenses in the survey sample for this RE-model divided by average observed expenses in the survey sample).
c. The merged survey sample contained many records with missing values for one of the pre-defined selective groups. We applied some corrections for the records with a missing value, which were due to some additional questions for only a specific group of respondents (e.g. limitations in performing daily activities for individuals older than 54 years). In addition, we excluded some records because respondents did not provide an answer (~3.3%). These corrections resulted into a survey sample that is used for quantifying the potential selection profits. A list of the corrections can be provided on request.
d. Observed expenses are annualized and weighted for the enrolment period in 2011. All expenses are rounded to the nearest Euro.
e. *Model 0*: constant (no risk equalization); *Model 1*: model 0 + age/gender (M = number of risk classes = 40); *Model 2*: model 1 + region (M = 50); *Model 3*: model 2 + source in income/age (M = 68); *Model 4*: model 3 + PCGs (M = 92); *Model 5*: model 4 + DCGs (M = 108); *Model 6*: model 5 + socioeconomic status/age (M = 120); *Model 7*: model 6 + MHC-groups (M = 127); *Model 8*: model 7 + DME-groups (M = 132).
f. PCG: Pharmaceutical Cost Group.
g. DCG: Diagnostic Cost Group. Individuals can be classified in only one DCG, the one with the highest follow-up costs.
h. MHC: Multiple-year High Cost-group.
i. DME: Durable Medical Equipment-group.

Further, average observed expenses in the survey sample are somewhat lower than those in the population: € 1,766 versus € 1,785 (not statistically different at 5%).

For our empirical application it is important to have no missing values for the questions that are used for defining the groups. However, many individuals did not respond to at least one of these questions. A main reason for this is that the survey includes additional questions for only individuals older than 12 years and some questions only for individuals older than 54 years, which were about the presence of chronic conditions or limitations in daily activities. Further, there were some follow-up questions. For all children and individuals younger than 54 years who had a missing value for these additional questions, we assumed that they do not have chronic conditions or limitations in performing daily activities. In addition, 537 records were excluded because these respondents did not provide an answer on one or more questions, while they belong to the targeted group for these questions (~3.3%). After these corrections, 15,604 complete records could be used for the analysis[15]. Of these individuals, 13,759 individuals were assigned to one or more pre-defined groups.

As a result of these corrections, average observed expenses reduced from € 1,766 to € 1,720 and the prevalence of a PCG, DCG, MHC-group, and DME-group reduced (Table 3.1). These statistics indicate that the excluded individuals had high above-average expenses and a below-average health status. Despite the exclusion of a small selective group, we can conclude that our survey sample after corrections can be considered reasonably representative for the Dutch population in terms of the prevalence of several risk factors, except for those individuals in nursing homes or other (mental care) institutions. In this sample, average age is 39 years, 16.5% of the individuals are classified into a PCG, 8.3% into a DCG, 5.6% into a MHC-group, and 0.9% to a DME-group. Combining these risk factors, 21.2% are classified to a PCG, DCG, MHC-group, or DME-group. These statistics do not largely deviate from those in the population (Table 3.1).

In terms of average expenses, there are some slight differences between the survey sample after corrections and the population. Mean total expenses in the survey sample are somewhat lower than those in the population: € 1,720 versus € 1,785 (not statistically different at 5%). Thus, the survey sample contains relatively somewhat healthier individuals in terms of average expenses than the Dutch population. To correct for these (small) differences in average expenses between the survey sample and the population, we raised individuals' residual expenses per RE-model by a factor equaling average RE-predicted expenses in the survey sample for this RE-model divided by average observed expenses in the survey sample. This way, average residual expenses per RE-model are zero in the survey sample, just as is the case in the population.

[15] A detailed description of the corrections for the records with a missing value for one or more of the pre-defined selective groups can be provided on request.

§ 3.3.5 Application of different methods

Since eight RE-models were tested according to three methods, our empirical analysis provided 24 estimates. All methods used a 5% significance level for testing the statistical significance of the criterion per iteration of the SR algorithm or the variables in the stepwise regression model. Further, since the pre-defined groups consisted of an over-representation of individuals with above-average expenses, the methods selected groups with positive average or total residual expenses or positive predicted residual expenses, resulting into defining the unfavorable group; all remaining individuals in the sample – those who were not selected by the SR algorithm but were assigned to one or more pre-defined groups plus those who were not assigned to any pre-defined group, or those with negative predicted residual expenses – form the complementary favorable group.

§ 3.4 EMPIRICAL FINDINGS

§ 3.4.1 Identification of the favorable and unfavorable group

Table 3.2 summarizes the definition of the favorable and unfavorable group according to each method that was examined per RE-model. This table clearly shows that the unfavorable group is a relatively small group with high above-average expenses. By definition, the complementary favorable group is a relatively large group with below-average observed expenses. The unfavorable group is approximately 23% to 42% of the total study population with average observed expenses of € 4,000 to € 2,800, depending on the method that is used for creating the group and the RE-model that is evaluated. Note that it is not useful to compare the unfavorable and favorable group across different methods for the same RE-model, because different pre-defined groups are identified.

For method 1 and 2 (i.e. the SR algorithm with average residual expenses and total residual expenses, respectively) it is clear which groups are statistically relevant after removing overlap, namely those groups that are selected by the algorithm to define the unfavorable group, which are 11 groups for RE-model 0 and 3 groups for RE-model 8 according to method 1; and 2 groups for RE-model 0 and 1 group for RE-models 1 to 8 according to method 2. Method 1 leads to selecting more groups than method 2, because relatively small groups can be selected by method 1, whereby the chance of overlap with other groups is smaller than first selecting relatively large groups. Examples of selected groups by method 1 are persons who contacted a home care practitioner in the past year, or those who contacted a home nurse. These groups are relatively small with high average observed expenses: 0.8% or 1.3% of the population with € 8,794 or € 9,065 expenses, respectively. These groups are not selected by method 2. According to this method, only relatively large groups are selected: under RE-model 0 the group of individuals who have a long-term disease (31.0% of the

Table 3.2: Descriptive statistics of the favorable and unfavorable group according to each of the three methods for creating mutually exclusive groups [a, b]

	Number of selected pre-defined groups	Unfavorable group		Favorable group	
		Percentage of study population	Average observed expenses in €'s	Percentage of study population	Average observed expenses in €'s
Method 1 [c,f]					
Model 0	11	29.4	3,644	70.6	918
Model 1	8	24.9	3,818	75.1	1,025
Model 2	8	24.9	3,818	75.1	1,025
Model 3	7	22.9	3,958	77.1	1,054
Model 4	4	38.4	3,042	61.6	897
Model 5	6	39.5	3,023	60.5	869
Model 6	6	39.5	3,023	60.5	869
Model 7	3	21.7	4,009	78.3	1,087
Model 8	3	21.7	4,009	78.3	1,087
Method 2 [d,f]					
Model 0	2	34.8	3,309	65.2	874
Model 1	1	37.4	3,045	62.6	929
Model 2	1	37.4	3,045	62.6	929
Model 3	1	37.4	3,045	62.6	929
Model 4	1	37.4	3,045	62.6	929
Model 5	1	37.4	3,045	62.6	929
Model 6	1	37.4	3,045	62.6	929
Model 7	1	37.4	3,045	62.6	929
Model 8	1	37.4	3,045	62.6	929
Method 3 [e,f]					
Model 0	n.a. [g]	33.0	3,535	67.0	828
Model 1	n.a. [g]	31.2	3,466	68.8	927
Model 2	n.a. [g]	31.2	3,464	68.8	928
Model 3	n.a. [g]	36.2	3,125	63.8	925
Model 4	n.a. [g]	35.9	2,953	64.1	1,029
Model 5	n.a. [g]	37.9	2,819	62.1	1,049
Model 6	n.a. [g]	37.9	2,819	62.1	1,049
Model 7	n.a. [g]	42.3	2,812	57.7	918
Model 8	n.a. [g]	42.2	2,809	57.8	926

Footnotes Table 3.2:

a. Statistics are based on survey sample after corrections.

b. *Percentage* in this table is based on the number of insured-years. The percentage individuals classified to unfavorable group plus the percentage individuals classified to the favorable group sum to 100% (= total study population; *N* (insured-years) = 15,535).

c. *Method 1*: stepwise removal algorithm based on average residual expenses as the criterion for determining the hierarchy of assigning individuals to only one pre-defined group of interest.

d. *Method 2*: stepwise removal algorithm based on total residual expenses as the criterion for determining the hierarchy of assigning individuals to only one pre-defined group of interest.

e. *Method 3*: stepwise regression analysis in order to predict residual expenses.

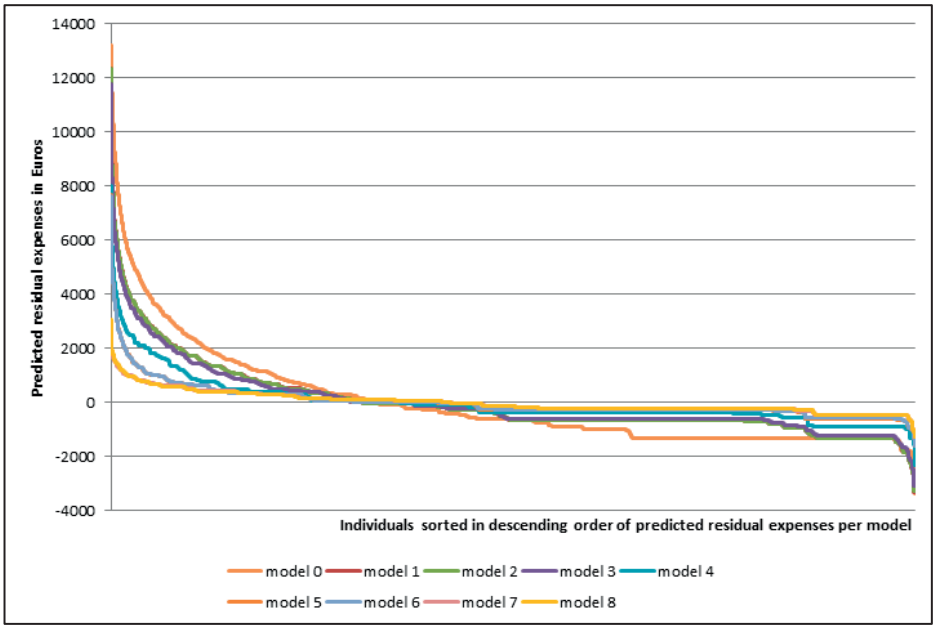
f. *Model 0*: constant (no risk equalization); *Model 1*: model 0 + age/gender (*M* = number of risk classes = 40); *Model 2*: model 1 + region (*M* = 50); *Model 3*: model 2 + source in income/age (*M* = 68); *Model 4*: model 3 + PCGs (*M* = 92); *Model 5*: model 4 + DCGs (*M* = 108); *Model 6*: model 5 + socioeconomic status/age (*M* = 120); *Model 7*: model 6 + MHC-groups (*M* = 127); *Model 8*: model 7 + DME-groups (*M* = 132).

g. n.a. = not applicable. For method 3, it is not possible to determine how many groups are selected after removing overlap between groups, because this method predicts residual expenses in order to create mutually exclusive groups.

population) and the groups of individuals who use complete dentures (9.6% of the population) are selected; and under RE-model 1 to 8 only the group of individuals who had contact with a medical specialist in the past year is selected: 37.5% of the population with average observed expenses of € 3,045. With method 3 (i.e. stepwise regression model) it is not possible to identify which groups are statistically relevant after removing overlap between these groups, because all relevant risk factors are used together for predicting residual expenses.

Figure 3.1 presents the predicted residual expenses for each RE-model based on method 3. This figure shows that the curve flattens as the predictive performance of the RE-model increases, indicating that the extra information has less explanatory power when the RE-model includes more risk adjusters and so, RE-residual expenses reduces. If we compare predicted residual expenses to RE-residual expenses for the unfavorable group under each of the RE-models, then we observe that average predicted residual expenses are larger than average RE-residual expenses for this group (statistics are not presented here): e.g. under RE-model 0, the insurer expects an under-prediction of € 2,026 per individual for the unfavorable group, while the RE-model under-predicts expenses for this group by € 1,815. This

Figure 3.1: The predicted residual expenses for all eight estimated RE-models [a,b]



Footnotes Figure 3.1:

a. Predicted residual expenses per RE-model based on method 3 (i.e. stepwise regression model) for creating mutually exclusive groups.

b. *Model 0*: constant (no risk equalization); *Model 1*: model 0 + age/gender (M = number of risk classes = 40); *Model 2*: model 1 + region (M = 50); *Model 3*: model 2 + source in income/age (M = 68); *Model 4*: model 3 + PCGs (M = 92); *Model 5*: model 4 + DCGs (M = 108); *Model 6*: model 5 + socioeconomic status/age (M = 120); *Model 7*: model 6 + MHC-groups (M = 127); *Model 8*: model 7 + DME-groups (M = 132).

implies that the usage of extra information enables the insurer to obtain a better prediction of individuals' profitability and so, to use this prediction to discriminate among the (very) unfavorable individuals and the (very) favorable individuals in the sample.

§ 3.4.2 Potential selection profits under various RE-models

Table 3.3 presents average residual expenses for 46 overlapping groups under eight RE-models. Table 3.4 provides the estimates of the potential selection profits under the same RE-models according to three methods for creating mutually exclusive groups. By definition, the estimates in absolute money amounts based on overlapping groups (Table 3.3) are higher than those based on mutually exclusive groups (Table 3.4) under the same RE-model, because individuals are counted multiple times in the calculation of the estimates for overlapping groups. Further, all groups are incorporated in the estimates in Table 3.3 even if groups are not statistically significant anymore when overlap between groups is removed. Comparing the percentage reduction in average residual expenses under each RE-model in Table 3.3 to the percentage reduction in potential selection profits under the same RE-models in Table 3.4 shows that aggregating residual expenses on overlapping

Table 3.3: Average residual expenses on 46 overlapping groups for eight RE-models [a]

	Average residual expenses, in €'s [b]	**Percentage reduction in average residual expenses** [c]
Model 0	1,462	-
Model 1	885	39.5
Model 2	878	39.9
Model 3	775	47.0
Model 4	493	66.3
Model 5	373	74.5
Model 6	373	74.5
Model 7	246	83.2
Model 8	242	83.4

Footnotes Table 3.3:

a. *Model 0*: constant (no risk equalization); *Model 1*: model 0 + age/gender (M = number of risk classes = 40); *Model 2*: model 1 + region (M = 50); *Model 3*: model 2 + source in income/age (M = 68); *Model 4*: model 3 + PCGs (M = 92); *Model 5*: model 4 + DCGs (M = 108); *Model 6*: model 5 + socioeconomic status/age (M = 120); *Model 7*: model 6 + MHC-groups (M = 127); *Model 8*: model 7 + DME-groups (M = 132).

b. Average residual expenses are calculated as: the sum of residual expenses over all individuals that are assigned to the 46 overlapping groups, divided by the total number of insured-years over all pre-defined groups; N(insured-years)= 61,041. The number of insured-years is larger than the total number of insured-years in the survey sample, because individuals can occur in multiple pre-defined groups.

c. Percentage reduction is calculated as: (profit model 0 minus profit of the model that is evaluated) divided by profit of model 0, multiplied by 100; e.g. (1,462-885)/1,462 = 39.5%.

groups *over*-estimates the percentage reduction in potential selection profits under a given RE-model[16]. However, the pattern in terms of the percentage reduction in potential selection profits across RE-models is similar: improving the RE-model leads to a reduction in potential selection profits, whereby Model 8 leads to the largest reduction of all RE-models. Therefore, using overlapping groups for estimating potential selection profits leads to an

Table 3.4: Potential selection profits under eight RE-models according to three methods for constructing mutually exclusive groups [a,b,c,d]

	Potential Selection Profit, in €'s [e,f]			**Percentage reduction in potential selection profits** [g]		
	Method 1	**Method 2**	**Method 3**	**Method 1**	**Method 2**	**Method 3**
Model 0	566	552	598	-	-	-
Model 1	371	364	404	34.5	34.1	32.4
Model 2	368	362	401	35.0	34.4	32.9
Model 3	309	339	356	45.4	38.6	40.5
Model 4	241	235	236	57.4	57.4	60.5
Model 5	160	154	166	71.7	72.1	72.2
Model 6	161	154	166	71.6	72.1	72.2
Model 7	98	124	135	82.7	77.5	77.4
Model 8	97	122	135	82.9	77.9	77.4
A perfect RE-model [h]	0	0	0	100	100	100

Footnotes Table 3.4:

a. *Model 0*: constant (no risk equalization); *Model 1*: model 0 + age/gender (*M* = number of risk classes = 40); *Model 2*: model 1 + region (*M* = 50); *Model 3*: model 2 + source in income/age (*M* = 68); *Model 4*: model 3 + PCGs (*M* = 92); *Model 5*: model 4 + DCGs (*M* = 108); *Model 6*: model 5 + socioeconomic status/age (*M* = 120); *Model 7*: model 6 + MHC-groups (*M* = 127); *Model 8*: model 7 + DME-groups (*M* = 132).

b. *Method 1*: stepwise removal algorithm based on average residual expenses as the criterion for determining the hierarchy of assigning individuals to only one pre-defined group of interest.

c. *Method 2*: stepwise removal algorithm based on total residual expenses as the criterion for determining the hierarchy of assigning individuals to only one pre-defined group of interest.

d. *Method 3*: stepwise regression analysis in order to predict residual expenses.

e. Potential selection profit is calculated as: the sum of residual expenses over all individuals that are assigned to the unfavorable group, divided by the total number of insured-years in the survey sample; *N*(insured-years)= 15,535. Note that the magnitude of the potential selection profit for the unfavorable group is equal to those of the complementary favorable group because both groups are complements and average residual expenses for the total population equals zero (because RE-model is estimated by Ordinary Least Squares).

f. All estimates of the potential selection profits are statistically significantly different from zero at 1%.

g. Percentage reduction is calculated as: (profit model 0 minus profit of the model that is evaluated) divided by profit of model 0, multiplied by 100%; e.g. (566-371)/566= 34.5%.

h. A perfect RE-model implies that this model adequately predicts expenses for all pre-defined groups of interest.

[16] Note that when the pre-defined set of groups contain an over-representation of individuals with below-average expenses, aggregating of residual expenses on these overlapping groups would result into an *under*-estimation of the potential selection profits when individuals with low residual expenses occur in multiple groups.

over-estimation of the reduction in potential selection profits under a given RE-model but it may not lead to another conclusion about the relative predictive performance of an RE-model, conditional on our set of pre-defined groups and the RE-models that are evaluated.

Table 3.4 shows that the potential selection profits in absolute money amounts differ across the methods for constructing mutually exclusive groups, given the same RE-model. Of all three methods, method 3 (i.e. the stepwise regression method) yields the largest potential selection profits under all RE-models that are evaluated, except for Model 4; however, the estimates for this model do not largely differ from each other: only € 5 difference between method 1 and 3. Further, the potential selection profits in absolute money amounts differ across the methods for Model 0 (the situation without RE), making it more useful to compare the percentage reduction in potential selection profits under a model compared to Model 0. Table 3.4 shows that the pattern in the percentage reduction in potential selection profits is similar across the methods: inclusion of more risk adjusters in the RE-model reduces the potential selection profits; e.g. under Model 8 the potential selection profits are reduced by 82.9%, 77.9%, and 77.4% for method 1, 2, and 3, respectively. Further, all RE-models are ranked in the same order of the predictive performance, implying that the method for constructing mutually exclusive groups does not lead to another conclusion about which RE-model has the highest model fit on the set of pre-defined groups of all RE-models that are evaluated. As expected, Model 8 consistently yields the largest reduction in the potential selection profits.

§ 3.5 CONCLUSIONS

Prior studies have evaluated RE-models on separate selective groups of interest in order to estimate the financial incentives for risk selection under these models. This study extends on this literature by estimating the potential selection profits on multiple groups simultaneously under a given RE-model. To perform such an evaluation it is important to avoid overlap between groups. This study develops and empirically applies three methods to create mutually exclusive groups from a set of overlapping selective groups of interest, with the purpose to estimate the potential selection profits under eight RE-models. These methods, a stepwise-removal algorithm with two variants and a stepwise regression model, make from multiple overlapping groups two distinctive groups: the favorable and the unfavorable group. For these groups, the total sum of residual expenses of a given RE-model provides an estimate of the potential selection profits for one year that under this RE-model can be exploited by attracting the total favorable group or rejecting the total unfavorable group, conditional on the set of pre-defined selective groups.

In our empirical application, aggregating average residual expenses for multiple overlapping groups over-estimates the potential selection profits under a given RE-model.

However, usage of overlapping groups instead of mutually exclusive groups did not lead to another conclusion about which RE-model leads to the largest percentage reduction in the potential selection profits and so, which model should be used in practice, conditional on our pre-defined set of groups and the RE-models that are evaluated. Further, our empirical application shows that the three methods for creating mutually exclusive groups lead to somewhat different estimates of the potential selection profits in absolute money amounts under a given RE-model. However, we find a consistent pattern in the percentage reduction in the potential selection profits under eight RE-models across the three methods for creating mutually exclusive groups. This implies that the usage of another method did not lead to another conclusion about which RE-model should be used in practice, conditional on our set of pre-defined groups and the RE-models that are evaluated.

The methods that are developed in this study can be a valuable instrument in practice for evaluating alternative RE-models on many different selective groups of interest, with the purpose to measure the financial incentives for risk selection under these RE-models. The methods applied here are broadly applicable to any RE-model and any type of information to create selective groups of interest. In some situations, information to create selective groups may not be routinely available. In these situations it may be worthwhile to invest in collecting this type of information; for example, by conducting health surveys.

§ 3.6 DISCUSSION

Our estimates indicate the *potential* selection profits under a given RE-model and *not* the actual selection profits that are expected by an insurer in practice. The methods in this study are developed with the purpose to use them for evaluating RE-models from a regulators' point of view. For this purpose, it is of interest which method yields the largest reduction in the potential selection profits, given a situation of an extreme form of asymmetric information whereby an insurer is able to exploit the extra information optimally. Here we assume that an insurer exploits health survey information on the individual level in order to define selective groups of interest and is able to act upon this information perfectly. Within this context, the estimates can serve as a basis for deciding which RE-model can best be used in practice for mitigating financial incentives for risk selection.

In practice, however, little is known about the type of information that is available to insurers and how and to what extent this information is exploited for engaging in risk selection, because often this is considered to be confidential (Breyer et al., 2012). The following five issues may be reasons why the selection profits that are expected by an insurer in practice may differ from our estimates based on an extreme form of asymmetric information and given the assumption that this information is used optimally. First of all, transaction costs from engaging in risk selection are not incorporated in our estimates. In practice,

insurers may ignore groups with small average residual expenses because of transaction costs and the statistical uncertainties about the netto profits of selection (van Barneveld et al., 2000). Second, insurers may have costs resulting from negative publicity when selecting specific groups. Third, we assume that an insurer is able to exploit the information on selective groups optimally and is able to act upon this information perfectly within one year. In practice, however, it may take several years to change the risk composition of the portfolio, because insurers may not be able to attract all favorable individuals or reject all unfavorable individuals within one year. Fourth, the potential selection profits are obtained by using information from one prior year. It may be more realistic, that insurers' planning horizon exceed one year and therefore, they may use multi-year information to identify selective groups with persistent high average residual expenses. In our estimates, we did not incorporate expected future profits due to unfavorable individuals becoming favorable individuals and vice versa (Beck & Zweifel, 1998; Beck et al., 2010; Welch, 1985). Fifth, an insurer may use other information than health survey information to obtain its own expectations about the profitability of specific groups, such as prior year's expenses. It is expected that the first three reasons may decrease the estimates of the selection profits, the fourth reason may increase the estimates of the selection profits, and for the fifth reason it is unclear how this will affect the estimates because that depends on the type of information that is used. Further research should investigate the netto effect of integrating all these issues in the estimates of the selection profits. If the netto effect of these issues is that insurers' estimates are lower than our estimates, the potential selection profits under an RE-model as estimated in this study do not have to equal zero in practice.

Our estimates of the potential selection profits use observed expenses as the reference point. In the RE literature, it has been suggested that measures-of-fit for evaluating RE-models should use another reference point than observed expenses because RE-models cannot, and do not have to, adjust for all differences in observed expenses; e.g. due to inefficiencies in the provision of care (van Veen et al., 2015a). To insurers, however, it does not matter whether variation in expenses is due to risk factors for which the regulator deems compensation appropriate or not. Consequently, the insurer may use any type of information to predict the profitability of specific groups of interest. In the extreme situation of a perfect RE-model that completely adjust for all cost variation related to risk factors for which the regulator deems compensation appropriate, there may still be potential selection profits for several selective groups that are of interest to the insurer; however, since these groups are defined by risk factors for which compensation is not desired, it is not of interest to improve the RE-model for these groups from a regulators' perspective. If, however, the selective groups are defined by risk factors for which the regulator deems compensation appropriate, the potential selection profits under a given RE-model are of interest to the regulator because insurers may have financial incentives to target risk selection on these

specific groups. It is worth noting that in practice it may not always be easy to purely distinguish groups for which the regulator desires compensation and those for which it is not.

Crucial is *which* groups are defined as a starting point for applying the methods as described in this study, because the estimates of the potential selection profits are conditional on this set of groups. From a regulators' perspective, this choice may be guided by several principles, including the questions whether it is profitable to select this group and whether selection actions can be targeted on this group. Prior studies particularly have focused on specific patient groups, groups with a poor general health status, or groups that have used certain healthcare facilities (e.g. Ash & Byrne-Logan, 1998; Ash et al., 2005; van Kleef et al., 2012a, 2012b, 2013b; Pope et al., 2000a, van Veen et al., 2015b). These groups are generally under-compensated by existing RE-models and so, they may be vulnerable to risk selection, whereby distortions of quality of care are a potential threat. In principle, any selective group can be defined that is of interest for evaluating an RE-model.

Since our estimates are conditional on the set of pre-defined groups and the eight RE-models that are evaluated, we cannot generally conclude to what extent a given RE-model provides financial incentives for risk selection in absolute money amounts and which method for creating mutually exclusive groups should be used in practice. The methods that are developed in this study are generally applicable for evaluating any RE-model, conditional on any set of pre-defined groups. The choice of the groups of interest and the RE-models that are evaluated may differ across regulators and countries. Given these choices, application of each of the three methods indicates which RE-model is the preferred model to be used in this situation.

Appendices Chapter 3

APPENDIX 3.1: DESCRIPTIVE STATISTICS

Table A.3.1: Descriptive statistics of the administrative dataset from the Dutch population of insured in 2011 (*N* = ~16.7 million), the two samples of this dataset used for estimation of the Dutch RE-model 2014, and the health survey sample used for the statistical analysis [a, b, c]

	Administrative data			Health survey data	
	Total dataset	**Training-sample**	**Validation-sample**	**Merged sample**	**Survey sample used for the analysis**
N(individuals)	16,688,961	8,327,580	8,361,381	16,141	15,604
N(insured-years) [d]	16,438,958	8,201,696	8,237,262	16,067	15,535
Mean total observed expenses, in €'s [e, f]	1,785 (5,978)	1,783 (5,944)	1,786 (6,011)	1,766 (5,364)	1,720 (5,325)
Median total observed expenses, in €'s	445	445	446	444	428
Mean predicted expenses per model, in €'s [e, f]					
Model 0			1,786 (0)	1,786 (0)	1,786 (0)
Model 1	n.a.	n.a.	1,785 (1,258)	1,752 (1,230)	1,708 (1,192)
Model 2	n.a.	n.a.	1,785 (1,264)	1,746 (1,237)	1,701 (1,199)
Model 3	n.a.	n.a.	1,784 (1,351)	1,721 (1,315)	1,673 (1,278)
Model 4	n.a.	n.a.	1,785 (2,023)	1,724 (1,968)	1,672 (1,918)
Model 5	n.a.	n.a.	1,786 (2,755)	1,749 (3,056)	1,698 (3,047)
Model 6	n.a.	n.a.	1,786 (2,756)	1,744 (3,053)	1,694 (3,045)
Model 7	n.a.	n.a.	1,787 (2,971)	1,762 (3,291)	1,712 (3,277)
Model 8	n.a.	n.a.	1,786 (2,975)	1,764 (3,297)	1,713 (3,283)
Age/gender					
Men 0-24 years	15.07%	15.09%	15.06%	15.38%	15.75%
Men 25-44 years	13.00%	13.01%	12.99%	11.37%	11.57%
Men 45-64 years	14.15%	14.13%	14.17%	14.50%	14.64%
Men 65-74 years	4.36%	4.35%	4.37%	4.70%	4.58%
Men ≥75 years	2.76%	2.76%	2.76%	2.66%	2.42%
Women 0-24 years	14.48%	14.50%	14.46%	15.70%	16.05%
Women 25-44 years	13.08%	13.08%	13.07%	12.56%	12.77%
Women 45-64 years	14.11%	14.09%	14.12%	14.86%	14.88%
Women 65-74 years	4.60%	4.59%	4.60%	4.65%	4.26%
Women ≥75 years	4.40%	4.39%	4.40%	3.62%	3.08%
Region					
Cluster 1-5	49.57%	49.60%	49.55%	47.23%	47.03%
Cluster 6-10	50.43%	50.40%	50.45%	52.77%	52.97%

Table A.3.1: (continued)

	Administrative data			Health survey data	
	Total dataset	**Training-sample**	**Validation sample**	**Merged sample**	**Survey sample used for the analysis**
Source of income					
Individuals <18 or >64 years	37.23%	37.26%	37.19%	39.35%	38.63%
Disability benefit	4.96%	4.96%	4.96%	4.10%	4.05%
Social security benefit	1.97%	1.97%	1.96%	1.27%	1.22%
Student	3.28%	3.27%	3.29%	3.36%	3.45%
Self-employed	4.05%	4.05%	4.06%	3.50%	3.56%
Others	48.51%	48.48%	48.54%	48.24%	49.09%
Socio-economic status					
Living on a home address with ≥15 persons	1.21%	1.22%	1.20%	0.34%	0.32%
Lowest income-class (deciles 1-3)	29.63%	29.65%	29.60%	26.91%	27.06%
Middle income-class (deciles 4-7)	39.52%	39.53%	39.52%	40.82%	40.50%
Highest income-class (deciles 8-10)	29.64%	29.60%	29.68%	31.94%	32.12%
Durable-medical equipment					
No equipment	99.19%	99.19%	99.19%	99.09%	99.14%
Insulin pump	0.11%	0.11%	0.11%	0.15%	0.16%
Catheter	0.39%	0.39%	0.39%	0.43%	0.40%
Colostomy	0.29%	0.29%	0.29%	0.31%	0.27%
Trachea-colostomy	0.02%	0.02%	0.02%	0.02%	0.03%
Multiple year high-costs					
No multiple year high costs	94.23%	94.23%	94.22%	94.12%	94.44%
2-years top 10%	1.00%	1.00%	1.00%	1.10%	1.07%
3-years top 15%	2.31%	2.31%	2.31%	2.31%	2.16%
3-years top 10%	1.06%	1.06%	1.06%	1.00%	0.96%
3-years top 7%	0.80%	0.79%	0.80%	0.80%	0.74%
3-years top 4%	0.46%	0.46%	0.46%	0.49%	0.45%
3-years top 1.5%	0.15%	0.15%	0.15%	0.17%	0.17%
% classified in one or more PCGs	17.30%	17.29%	17.31%	17.22%	16.53%
% classified in multiple PCGs	3.47%	3.47%	3.47%	3.46%	3.17%

Table A.3.1: (continued)

	Administrative data			Health survey data	
	Total dataset	**Training-sample**	**Validation sample**	**Merged sample**	**Survey sample used for the analysis**
% classified in a DCG	8.65%	8.64%	8.67%	8.72%	8.30%
Combinations of risk classes					
% classified into a PCG, DCG, DME-group, and/or MHC-group	22.05%	22.03%	22.06%	22.02%	21.24%
% *not* classified into a PCG, DCG, DME-group, and MHC-group	77.95%	77.97%	77.94%	77.98%	78.76%

Footnotes Table A.3.1:

a. The estimation- and validation-sample were used to estimate the RE-models. The survey sample (after correction for the missing values) – the last column – was used for estimating the potential selection profits of RE-models.

b. RE-predicted expenses were not available for the total administrative dataset and the estimation-sample of this administrative dataset, because the coefficients of the RE-models were estimated on the estimation-sample and expenses were predicted on the validation-sample. The models were estimated by Ordinary Least Squares (OLS). Average predicted expenses by some models (slightly) deviated from average observed expenses in the validation sample, because of the use of a split sampling approach. Further, as a result of merging the validation sample to the health survey, average predicted expenses in the survey sample deviated from average predicted expenses in the validation-sample. For examining the residuals on the groups in this survey sample, the residuals were calibrated in such a way that average residual for each model on the survey sample equals zero. This was done to maintain the zero-sum principle of RE-models, implying that the magnitude of the potential selection profit on the unfavorable group equals the magnitude of the potential selection profits for the complementary favorable group (though an opposite sign).

c. *Model 0*: constant (no risk equalization); *Model 1*: model 0 + age/gender (M = number of risk classes = 40); *Model 2*: model 1 + region (M = 50); *Model 3*: model 2 + source in income/age (M = 68); *Model 4*: model 3 + PCGs (M = 92); *Model 5*: model 4 + DCGs (M = 108); *Model 6*: model 5 + socioeconomic status/age (M = 120); *Model 7*: model 6 + MHC-groups (M = 127); *Model 8*: model 7 + DME-groups (M = 132).

d. N(insured-years) 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 enrolled the full year.

e. Expenses are annualized and weighted for the enrolment period. All expenses are rounded to the nearest Euro.

f. Standard deviation is presented in parentheses. The sum of the weights minus one is used as the variance divisor.

APPENDIX 3.2: MODELS' PREDICTIVE PERFORMANCE ON THE FULL SAMPLE

Table A.3.2: The R-squared (R^2), Cumming's Prediction Measure (CPM), and Mean Absolute Prediction Error (MAPE) for the estimated RE-models [a,b]

	R^2 (in %) [c]	CPM (in %) [d]	MAPE (in €'s) [e]
Model 0	0	0	2,090
Model 1	4.4	9.5	1,892
Model 2	4.4	9.6	1,890
Model 3	5.1	10.8	1,864
Model 4	11.3	17.2	1,731
Model 5	20.7	21.4	1,642
Model 6	20.8	21.5	1,642
Model 7	24.2	24.9	1,570
Model 8	24.3	25.0	1,568

Footnotes Table A.3.2:

a. The models were evaluated on the validation-sample of the administrative dataset (N = 8,361,381).

b. *Model 0*: constant (no risk equalization); *Model 1*: model 0 + age/gender (M = number of risk classes = 40); *Model 2*: model 1 + region (M = 50); *Model 3*: model 2 + source in income/age (M = 68); *Model 4*: model 3 + PCGs (M = 92); *Model 5*: model 4 + DCGs (M = 108); *Model 6*: model 5 + socioeconomic status/age (M = 120); *Model 7*: model 6 + MHC-groups (M = 127); *Model 8*: model 7 + DME-groups (M = 132).

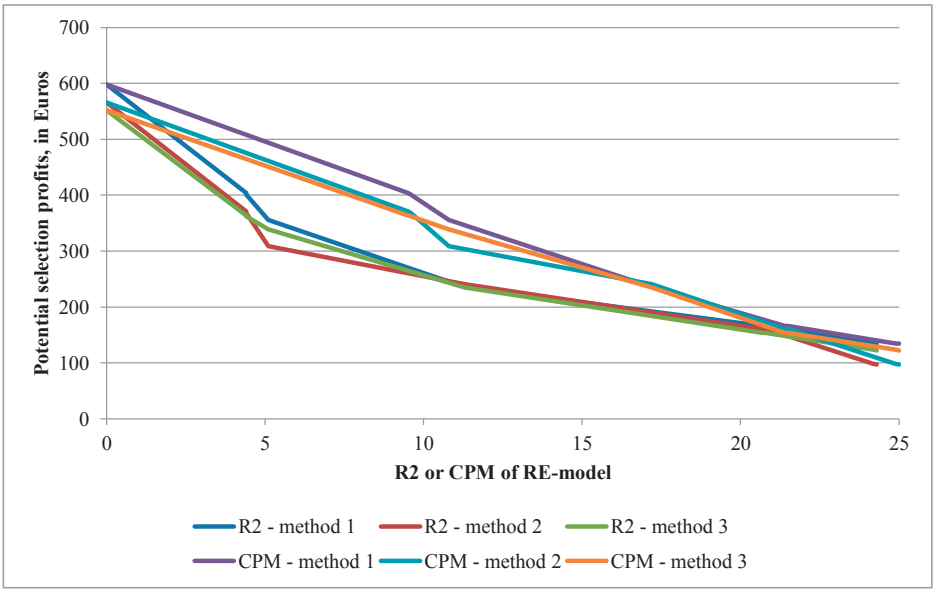
c. R^2 was calculated as one minus the (sum of squared residuals divided by the total sum of squares).

d. CPM was calculated as the mean absolute residual divided by the mean absolute deviation of observed expenses to average observed expenses.

e. MAPE was calculated as the mean absolute deviation of observed expenses to predicted expenses.

APPENDIX 3.3: RELATIONSHIP BETWEEN THE R^2 AND CPM

Figure A.3.3: Relationship between the R^2 and CPM of the RE-model on the full sample and the potential selection profits for multiple groups simultaneously [a-h]



Footnotes Figure A.3.3:

a. *Method 1*: stepwise removal algorithm based on average residual expenses as the criterion for determining the hierarchy of assigning individuals to only one pre-defined group of interest.
b. *Method 2*: stepwise removal algorithm based on total residual expenses as the criterion for determining the hierarchy of assigning individuals to only one pre-defined group of interest.
c. *Method 3*: stepwise regression analysis in order to predict residual expenses.
d. Potential selection profit is calculated as: the sum of residual expenses over all individuals that are assigned to the unfavorable group, divided by the total number of insured-years in the survey sample; N(insured-years)= 15,535. Note that the magnitude of the potential selection profit for the unfavorable group is equal to those of the complementary favorable group because both groups are complements and average residual expenses for the total population equals zero (RE-model is estimated by Ordinary Least Squares).
e. *Model 0*: constant (no risk equalization); *Model 1*: model 0 + age/gender (M = number of risk classes = 40); *Model 2*: model 1 + region (M = 50); *Model 3*: model 2 + source in income/age (M = 68); *Model 4*: model 3 + PCGs (M = 92); *Model 5*: model 4 + DCGs (M = 108); *Model 6*: model 5 + socioeconomic status/age (M = 120); *Model 7*: model 6 + MHC-groups (M = 127); *Model 8*: model 7 + DME-groups (M = 132).
f. R^2: R-squared, calculated as one minus the (sum of squared residuals divided by the total sum of squares). This R^2 is adjusted for the number of variables included in the RE-model.
g. CPM: Cumming's Prediction Measure, calculated as the mean absolute residual divided by the mean absolute deviation of observed expenses to average observed expenses.
h. Note that the MAPE can also be plotted against the potential selection profits; however, this is just as the CPM a linear measure. Further, since the MAPE is not standardized it is easier to present the R^2 and CPM in one figure. The CPM is the linear counterpart of the R^2.

Part II

Improving the Predictive Performance of Risk Equalization Models

Chapter 4

Cost and Diagnostic Information from Multiple Prior Years

ABSTRACT

Currently-used risk equalization models do not adequately compensate insurers for predictable differences in individuals' healthcare expenses. Consequently, insurers face incentives for risk rating and risk selection, both of which jeopardize affordability of coverage, accessibility to healthcare, 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 to 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.

§ 4.1 INTRODUCTION

§ 4.1.1 Background

Several countries world-wide have implemented risk equalization (RE) into their (competitive) health insurance scheme. RE is a system of prospective risk-adjusted payments to compensate health insurers or health plans for predictable differences in individuals' healthcare expenses. The principal goals of RE are (i) to achieve affordability of health insurance for high-risk individuals and (ii) to mitigate financial incentives for insurers to engage in risk selection (van de Ven & Ellis, 2000). The latter is particularly relevant for competitive health insurance schemes with premium regulation as found in Belgium, Germany, Israel, the Netherlands, and Switzerland.

Schokkaert & van de Voorde (2004, 2006, 2009) have advocated that the calculation of the risk-adjusted payments involves two steps. The first step purely focuses on the estimation of the prediction model, with the aim to explain variation in individual healthcare expenses and obtain accurate predictions, as much as possible. Schokkaert & van de Voorde (2004, 2006, 2009) 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. In the second step, the estimated model is used for calculating the 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 (Schokkaert & van de Voorde 2004, 2006, 2009). 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 purely focuses on the first step of the calculation of risk-adjusted payments; i.e. on the estimation of the prediction model.

§ 4.1.2 Development of risk equalization models

Over the past two decades, the predictive performance of the models used in RE has substantially improved as a result of the development of diagnostic-based and pharmacy-based risk adjusters (Adams et al., 2002; Fishman et al., 2003; Fleishman et al., 2006; Gilmer et al., 2001; Hughes et al., 2004; Kronick et al., 2000; Lamers, 2001; Lamers & van Vliet, 2003, 2004; Pope et al., 2000a; Prinsze & van Vliet, 2003), with over the past five years an increasing attention in the RE literature on the development of indicators for health status based on prior utilization or costs (e.g. van Kleef & van Vliet, 2010, 2012), and risk adjusters based on self-reported health or chronic conditions (e.g. DeSlavo et al., 2009; Fleishman et al., 2006; Stam et al., 2010b). Examples of diagnostic-based and pharmacy-based models are

those used in Belgium, Germany, the Netherlands, and the U.S.. Several studies, however, have shown that even these sophisticated models do not adequately predict individual expenses, especially for high-risk individuals (Barry et al., 2012; Behrend et al., 2007; van Kleef et al., 2012a, 2012b). 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 healthcare, and quality of healthcare (van de Ven & Ellis, 2000; van de Ven & Schut, 2011). For example, insurers can select risks by offering less attractive benefits, or not contracting high-quality care, or providing poor services to high-risk groups (Newhouse, 1996; van de Ven & Ellis, 2000). 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.

§ 4.1.3 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' (van Kleef & van Vliet, 2012; van Kleef et al., 2012b). 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 (Lamers & van Vliet, 1996; Lamers, 1997; Stam & van de Ven, 2008; van Kleef & van Vliet, 2012; van Kleef et al., 2012b). 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 (N = 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 statistically significantly contribute 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 (van de Ven & Ellis, 2000). 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 does 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 healthcare 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 only (very) limitedly applied in practice for calculating the risk-adjusted payments, because risk adjusters based on prior costs and/or prior utilization may reduce incentives for efficiency (e.g. Lamers & van Vliet, 1996; Lamers, 1997; van Vliet & van de Ven, 1992, 1993). Following the approach of Schokkaert & van de Voorde (2004, 2006), we do not have to be concerned with these normative choices about C- and R-type risk factors in our empirical analysis, because we purely focus 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 policymakers in countries with a RE scheme or for those who want to implement RE into their health insurance scheme. Although this

study uses administrative data from the Netherlands, regulators and policymakers 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 chapter 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 into their health insurance scheme.

§ 4.2 DATA AND METHODS

§ 4.2.1 Administrative data and health survey data

Two datasets are used for the empirical analysis. The first dataset contained individual-level administrative data for the Dutch population in the period 2006 to 2009. The sample analyzed in this study consisted of individuals who were enrolled, for a part or a full year, in each of the four 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 healthcare 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 three months in year *t*

[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 new borns 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 four years due to switching behaviour of individuals in prior years. Records of individuals who did not switch in year *t*, but who switched in one or more of the three prior years were copied (duplicates) when merging the administrative data of four years. These duplicate records were weighted by a value of 0.5 in the estimation of the model. There were no individuals who switched more than once of insurer during one year (which would mean that more than two records would be generated during the merging process).

and had € 100 expenses was given a weight of 0.25 and € 400 annual expenses. By applying this method, mean predicted expenses in year *t* equals mean observed expenses in year *t*. Table 4.1 shows some descriptive statistics. Mean total expenses in year *t*, *t*-1, *t*-2, and *t*-3 are € 1,689, € 1,639, € 1,495, and € 1,383 respectively. In the study population in year *t*, average age is 41.5 years, 2.8% of the individuals are 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. The present study merged the administrative dataset with the survey data at the individual level using an anonymous, unique identification variable (N = 7,979)[4]. The health status information was used to define groups in the population to assess the predictive performance at the group level. Given the administrative data and the health

Table 4.1: 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 four-year period (N = 13.8 million)

	Mean (std.)[d]
Total observed expenses (in €'s)[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
Proportion classified into a DCG[c] in year *t*-1	0.028
Proportion classified into a PCG in year *t*-1	0.177
Proportion classified into more than one PCG in year *t*-1	0.035

Footnotes Table 4.1:

a. The expenses in year *t* were annualized and weighted for the enrolment period. The expenses in the year *t*-1, *t*-2, and *t*-3 refer to observed expenses. All expenses were rounded to the nearest euro.
b. The prediction year t is the year 2009. Year *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.
d. Std. = standard deviation.

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

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

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.

§ 4.2.2 Model estimation

Model 1 – 4: proxies for currently-used models

As a first step, four models are estimated to compare the outcomes of the two newly-developed models to those models. All independent variables in these models are dummy variables defining different risk classes in the population. Model 1 only includes an intercept 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 4.1 describes the specification of these variables. A more detailed description is well-documented elsewhere (van Kleef & van Vliet, 2010). 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 three prior years (*M* = 119). Table 4.2 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,

Table 4.2: Description of the independent variables for each estimated model

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 + 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	113
Model 4 (Dutch model of 2012)	variables of Model 3 + 6 dummy variables for multi-year high costs risk classes	119
Model 5 (multi-year health-based model)	variables of Model 3 + 26 dummy variables for DCGs in year *t*-2 and *t*-3[a] + 40 dummy variables for PCGs in year *t*-2 and *t*-3[b]	179
Model 6 (multi-year health/cost-based model)	variables of Model 5 + 694 dummy variables for percentiles of ten types of expenses from year *t*-1, *t*-2, and *t*-3[c,d] + 30 continuous variables for ten types of expenses from year *t*-1, *t*-2, and *t*-3	903[e]

Footnotes Table 4.2 (Appendix 4.1 gives a more detailed description of the variables in Models 1 to 4):

a. For each year, we had 13 Diagnostic Cost Groups.

b. For each year, we had 20 Pharmaceutical Cost Groups.

c. 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.

d. 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 respectively 225, 235, and 234 dummy variables for percentiles of expenses.

e. 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 (~62%), using a 5% significance level.

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 5 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 one year increases the likelihood of having high expenses in the next year(s) (Garber et al., 1998; Monheit, 2003). All individuals with zero expenses per type of expenses were classified into a 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 (Fox, 2008; Pindyck & Rubinfeld, 1998; Thompson, 1978). 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 primarily focus 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 (Fox, 2008; Pindyck & Rubinfeld, 1998). For purposes of prediction, however, multicollinearity is not of particular interest, because we are only interested in the predictive power of the model and not so much 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 (Babyak, 2004; McIntyre et al., 1983). Therefore, the total sample was split into a training and validation sample. In the fourth step of the analysis, administrative data is merged with

[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.

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 for the Dutch population that was enrolled during the study period (Table 4.3). All six models examined in this study were estimated on the training sample and the coefficients of the variables in these models are used to predict individual expenses in the validation sample[7].

All six models were assumed to be linear in the coefficients and included an intercept. The use of Ordinary Least Squares-models (OLS) on untransformed data for predicting individual expenses has been widely discussed in literature, because OLS may not fit the distributional properties of healthcare expenses very well (Basu & Manning, 2009; Buntin & Zaslavsky, 2004; Manning & Mullahy, 2005; Manning et al., 2005; Veazie et al., 2003). 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 policymakers and therefore, OLS on untransformed data has been widely adopted 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 (Dunn et al., 2003; Dunn, 2003; Jones, 2010; Mihaylova et al., 2011; Powers et al., 2005; van Vliet & van de Ven, 1993). Therefore, we expect that we would have found quite similar results with other estimation methods than OLS.

§ 4.2.3 Model evaluation

As a fourth step, the predictive performance of the estimated models was assessed and compared at both the population and group level. By doing so, it is possible to examine how well the models predict expenses for the total sample and for specific groups 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.

[7] Model parameters of each estimated model can be provided on request.

Table 4.3: 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[a]	Administrative data[b]			Survey data[e]
	Total sample	Training-sample[c]	Validation-sample[d]	
N(records)	14,001,206	6,999,827	7,001,379	8,091
N(individuals)	13,801,415	6,900,221	6,901,194	7,979
N(insured-years)[f]	13,712,676	6,855,800	6,856,876	7,938
Expenses				
Mean observed expenses in €'s	1,689	1,688	1,689	1,706
Mean predicted expenses Model 1 in €'s	n.a.	n.a.	1,689	1,689
Age/gender				
Men 0-24 years	13.81%	13.80%	13.81%	15.67%
Men 25-44 years	13.41%	13.41%	13.41%	11.64%
Men 45-64 years	14.11%	14.10%	14.11%	13.57%
Men 65-74 years	4.38%	4.38%	4.39%	4.64%
Men >75 years	2.96%	2.96%	2.96%	2.95%
Women 0-24 years	13.29%	13.28%	13.30%	14.49%
Women 25-44 years	13.76%	13.78%	13.73%	12.68%
Women 45-64 years	14.59%	14.58%	14.60%	14.73%
Women 65-74 years	4.78%	4.78%	4.77%	5.00%
Women >75 years	4.93%	4.93%	4.93%	4.64%
Region				
Cluster 1-5	50.18%	50.19%	50.18%	47.55%
Cluster 6-10	49.82%	49.81%	49.82%	52.45%
Source of income				
Individuals <18 years or >64 years	35.61%	35.62%	35.60%	39.53%
Disability benefit	5.36%	5.36%	5.36%	4.85%
Social security benefit	2.01%	2.01%	2.00%	1.18%
Self-employed	4.15%	4.16%	4.15%	3.65%
Others	52.87%	52.85%	52.88%	50.80%
Socio-economic status				
Living on a home address with ≥15 persons	1.40%	1.41%	1.39%	0.38%
Lowest income-class (deciles 1-3)	29.50%	29.51%	29.49%	29.01%
Middle income-class (deciles 4-7)	40.21%	40.19%	40.24%	41.07%
Highest income-class (deciles 8-10)	28.89%	28.89%	28.88%	29.54%
% classified in one or more PCGs	17.70%	17.68%	17.72%	17.83%
% classified in multiple PCGs	3.54%	3.53%	3.56%	3.42%
% classified in a DCG	2.82%	2.81%	2.82%	2.64%

Footnotes Table 4.3:

a. Prediction year *t* is 2009.
b. Individual-level administrative data from 2006 to 2009 is used.
c. The models are estimated on this sample.
d. Expenses of individuals are predicted on this sample.
e. Models' predictive performance at the group level is assessed on this sample. The health survey is conducted in year *t*-1; i.e. 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.

Models' predictive performance at the group 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 groups defined by information from the training sample than information from the validation sample and on groups matching (or highly correlated with) the risk cells of the model (Cumming et al., 2002). 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 groups can provide a good indication of the extent to which models compensate insurers for differences in expenses between groups. This method is also applied in other studies (Stam & van de Ven, 2006, 2008; van Kleef et al., 2012a, 2012b).

General demographic risk characteristics in the dataset used for the model evaluation at the group level are comparable to those of the training sample and validation sample, providing evidence for the representativeness of the health survey respondents for the Dutch population (Table 4.3). 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 on 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 main reason for the latter is that the health survey is mainly targeted on 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 group of institutionalized individuals[8].

Specifically, information on self-reported health status, (long-term) diseases and conditions, and healthcare utilization is used to construct forty-five groups. These groups are defined in such a way that they include a relatively large proportion of high-risk individuals (e.g. chronically ill). These groups are comparable to those defined by van Kleef and colleagues and Stam & van de Ven (van Kleef et al., 2012a, 2012b; Stam & van de Ven, 2006). The groups are 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 groups are defined by 'yes/no'-questions. Table A.2 describes the definition of the evaluation-groups (see page 257).

A (two-sided) T-test is applied to test whether the MPEs on groups 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 4.3 shows that mean total observed expenses differs from mean total predicted expenses of Model 1 in

[8] Based on an empirical analysis of Dutch administrative data from 2007, under-predictions varying from € 300 up to € 1,400 can be expected on groups with a relatively large proportion of institutionalized individuals (Schut & van de Ven, 2010).

the survey sample. Therefore, the MPEs for each model in the survey sample were calibrated as follows: individual observed expenses were raised by a factor equaling average predicted expenses in the survey sample divided by average observed expenses in the survey sample. These calibrated MPEs were used to assess models' predictive performance on groups and to test the statistical significance of the MPEs.

§ 4.3 RESULTS

§ 4.3.1 Predictive performance at the population level

The results in Table 4.4 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. 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. 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. The R^2-value of Model 5 is 24.84% and the MAPE-value is € 1,537, 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

Table 4.4: Adjusted-R^2 and MAPE of the estimated models

	Adjusted R^2 (in %)[a]	MAPE[b] (in €'s)
Model 1 (no risk equalization)	0.00	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

Footnotes Table 4.4:

a. 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). 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. MAPE = Mean Absolute Prediction Error, which was calculated as: the average of the absolute differences between predicted expenses and observed expenses.

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 (Ash et al., 1989; Lamers & van Vliet, 1996; van Vliet & van de Ven, 1993), these results confirm the predictive power of cost and diagnostic information from multiple prior years.

§ 4.3.2 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) (Fox, 2008; Thompson, 1978). 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 to € 1,349.06, with the MAPE-value of Model 6 being € 1,348.96. These results indicate the robustness of the specification of Model 6 as applied here for predicting individual expenses.

§ 4.3.3 Predictive performance at the group level

Based on analyzing the MPE-values of all models for the forty-five groups, for fourteen groups 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 4.5). For seven groups all estimated models have produced statistically significant MPE-values, implying that adding risk adjusters based on cost and diagnostic information from three

Table 4.5: Groups for which the Mean Prediction Error in year *t* is *NOT statistically significantly* different from zero for Model 6 [a,b,c]

Groups (based on health survey data from year *t*-1)	**Size, in %**	**Mean observed expenses year *t*, in €'s**	**MPE in year *t*, in €'s**					
			Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
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								
Hearing-aid (age ≥ 4 years)	3.6	5,017	-3,328***	-1,221**	-1,116**	-835*	-1,089**	-604
Durable medical equipment (age ≥ 12 years)	5.8	5,279	-3,590***	-2,292***	-1,451***	-966**	-1,374***	-603

Footnotes Table 4.5:

a. MPE = Mean Prediction Error, is calculated as mean of (predicted expenses minus observed expenses).

b. ***: Statistically significantly different from zero with a *P*-value ≤ 0.01; **: Statistically significantly different from zero with a *P*-value ≤ 0.05; *: Statistically significantly different from zero with a *P*-value ≤ 0.10 (based on an one-sample two-sided *T*-test).

c. In this study, the prediction year *t* is 2009. The column of total expenses presents the calibrated total expenses. Total expenses and predicted expenses in the sample with health survey information were calibrated 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).

Table 4.6: Groups for which the Mean Prediction Error in year *t* is *statistically significantly* different from zero for Model 6 [a,b,c]

Groups (based on health survey data from year *t*-1)	**Size, in %**	**Mean observed expenses year *t*, in €'s**	**MPE in year *t* in €'s**					
			Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
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*

Footnotes Table 4.6:

a. MPE = Mean Prediction Error, is calculated as mean of (predicted expenses minus observed expenses).

b. ***: Statistically significantly different from zero with a *P*-value ≤ 0.01; **: Statistically significantly different from zero with a *P*-value ≤ 0.05; *: Statistically significantly different from zero with a *P*-value ≤ 0.10 (based on an one-sample two-sided *T*-test).

c. In this study, the prediction year *t* is 2009. The column of total expenses presents the calibrated total expenses. Total expenses and predicted expenses in the sample with health survey information were calibrated 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).

prior years is not sufficient to adequately predict expenses for these groups (Table 4.6). Finally, for twenty-four groups 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 4.2). In the remainder of this section, we purely focus on the first two types of results; i.e. on Tables 4.5 and 4.6.

For all pre-defined groups expenses in year *t* are (far) above average expenses in the total sample in year *t*, indicating that all groups contain (as expected) a relatively high proportion of high-risk individuals. Further, for most groups the MPE has a negative value, which means that the models under-predict expenses for these groups. These under-predictions indicate that expenses for the complementary groups (i.e. the low-risk individuals) are over-predicted. Notice that positive MPE-values imply that the model over-predicts expenses for this group. When interpreting the results in Tables 4.5 and 4.6, it should be taken into consideration that the same individual may occur in multiple subgroups.

The results in Table 4.5 show that models with more risk adjusters produce more accurate predictions at the group level than models using less risk adjusters. For example, Model 1 in Table 4.5 shows substantially negative MPE-values for all groups, 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 groups, 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 groups: e.g. for individuals with OECD limitations in moving (age ≥ 12 years), individuals with a low score on the SF-12 scales (age ≥ 12 years), individuals with limitations in daily activities (age ≥ 55 years), or individuals who reported two or more diseases (age ≥ 12 years). Model 4, however, further improves the performance for all groups in Table 4.5, which is due to the inclusion of a risk adjuster for 'multiple-year high costs'. Further, Model 6 outperforms all other models on all groups in Table 4.5. The MPE-values on all groups in Table 4.5 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 groups in the population.

However, the results in Table 4.6 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 groups. Under-predictions (statistically significantly different from zero) remain for

individuals who reported a poor general health status (age ≥ 12 years), one or more long-term diseases (age ≥ 12 years), a myocardial infarction or other serious heart disease (age ≥ 12 years), psoriasis (age ≥ 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 ≥ 12 years), three or more self-reported diseases or disorders (age ≥ 12 years), or use of complete dentures (age ≥ 16 years). Apparently, these groups 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.

§ 4.4 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 with 8 percentages points at the population level. At the group 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' healthcare expenses, implying that there are still under-predictions (that are statistically significantly different from zero) for certain high-risk groups 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, our findings indicate that financial incentives for risk rating and/or risk selection can be substantially reduced by using cost and diagnostic information from multiple prior years, but even using this information does not remove these incentives completely.

§ 4.5 DISCUSSION

§ 4.5.1 Methodological limitations and points for further research

The empirical analysis and the data used to illustrate the potential for improving the predictive performance of 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 (Garber et al., 1998; Lamers, 1997; Monheit, 2003). 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 (van Kleef et al., 2012c). 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. Basu & Manning, 2009; Buntin & Zaslavsky, 2004; Duan et al., 1983; Manning & Mullahy, 2001; Manning et al., 2005; Veazie et al., 2003). 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.

§ 4.5.2 Health-policy implications

As Schokkaert & van de Voorde have advocated, the calculation of the risk-adjusted payments used in practice involves two steps (Schokkaert & van de Voorde, 2004, 2006, 2009). In the first step, the model is estimated with the aim to explain variation in individual healthcare expenses and obtain accurate predictions, as much as possible. The second step uses the estimated model to calculate the risk-adjusted payments, which involves normative choices by the regulator on 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 on the extent to which currently-used RE-models can be improved in practice. Our findings should be interpreted while bearing in mind the following.

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, the C-type risk factors, and risk factors for which the regulator does not desire compensation, the R-type risk factors (Schokkaert & van de Voorde, 2006). According to the approach of Schokkaert & van de Voorde, 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 (Schokkaert & van de Voorde, 2004, 2006, 2009). In the second step, the effects of these R-type risk factors can be neutralized by using e.g. the average value of this risk factor or to use the same value for all individuals in the population. Following this approach, regulators could use the developed models 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 so, to derive the risk-adjusted payments used in practice. Note that the choice about 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 policymakers could relatively easily improve the predictive performance of currently-used models by including cost and diagnostic information from multiple prior years.

To conclude, currently-used RE-models do not adequately compensate insurers for predictable differences in individuals' healthcare expenses, which faces insurers with incentives for risk rating and risk selection, both of which jeopardize affordability of coverage, accessibility of healthcare, 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. 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.

Appendices
Chapter 4

APPENDIX 4.1: DEFINITION OF THE RISK ADJUSTERS

Table A.4.1: Definition of risk adjusters included in estimated RE-models based on administrative data from 2009

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 to 4 years, 5 to 9 years, 10 to 14 year, 15 to 17 years, 18 to 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 to 34 years, 34 to 44 years, 45 to 54 years and 55 to 64 years). There is a separate risk class for individuals younger than 14 years or older than 64 years old.	16
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 to 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, two years in top 15% of total expenses and a separate class for those individuals who do not have high expenses in multiple years.	6

Footnotes Table A.4.1:

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 4.2: THE MEAN PREDICTION ERROR FOR SOME EVALUATION-GROUPS

Table A.4.2: Groups for which the Mean Prediction Error in year *t* was already *NOT statistically significantly* different from zero for Model 1, 2, 3, or 4 [a,b,c]

Groups (based on health survey data from year *t*-1)	**Size, in %**	**Mean observed expenses in year *t*, in €'s**	**MPE in year *t*, in €'s**					
			Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
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
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
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

Table A.4.2: (continued)

Groups (based on health survey data from year *t*-1)	Size, in %	Mean observed expenses in year t, in €'s	MPE in year t, in €'s					
			Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
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

Footnotes Table A.4.2:

a. MPE = Mean Prediction Error, is calculated as mean of (predicted expenses minus observed expenses).

b. ***: Statistically significantly different from zero with a *P*-value ≤ 0.01; **: Statistically significantly different from zero with a *P*-value ≤ 0.05; *: Statistically significantly different from zero with a *P*-value ≤ 0.10 (based on an one-sample two-sided *T*-test).

c. In this study, the prediction year *t* is 2009. The column of total expenses presents the calibrated total expenses. Total expenses and predicted expenses in the sample with health survey information were calibrated 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).

Chapter 5

Interaction Terms between Existing Risk Adjusters

ABSTRACT

This chapter 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 identified interaction terms were used as additional risk adjusters in the RE-model. We found evidence that interaction terms can improve the predictive performance of the RE-model. Because regression trees are not robust, additional criteria have to be used 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. Our analysis shows that interaction terms can reduce financial incentives for risk selection but cannot eliminate them.

§ 5.1 INTRODUCTION

§ 5.1.1 Background

Risk equalization (RE) models are widely used for calculating risk-adjusted payments to health insurers in order to compensate them for predictable differences in individuals' healthcare expenses. Belgium, Germany, Israel, the Netherlands, Switzerland, and U.S. (Medicare) have used RE-models for several decades (Ash et al., 1989; van de Ven et al., 2007). Since January 2014, RE is used in the health insurance exchanges in the U.S. (Kautter et al., 2014). 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.

The RE literature during the past three decades has been largely devoted to investigating the predictive power of new types of risk adjusters in RE-models; e.g. demographic risk adjusters, morbidity-based risk adjusters relying on pharmaceutical or diagnostic information, or cost-based risk adjusters (e.g. Adams et al., 2002; Fishman et al., 2003; Hughes et al., 2004; Kronick et al., 2000; Lamers & van Vliet, 2003; Pope et al., 2000a). 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 healthcare expenses. To a certain extent this risk heterogeneity may be explained by interactions between the risk classes in the RE-model. Some studies have demonstrated that interactions can improve model's predictive performance (Buchner et al., 2014; Pope et al., 2000b; 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, have included some first-order interactions; e.g. 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 useful to adequately predict healthcare expenses for selected non-random groups, such as individuals with comorbidities (Pope et al., 2004). Several studies have shown that even the morbidity-based RE-models do not adequately predict healthcare expenses (e.g. Behrend et al., 2007; van Kleef et al., 2014; Payne et al., 2000). In the presence of premium regulation, under- and over-predictions of expenses for selected non-random groups provide health insurers with financial incentives for risk selection, which is a potential threat to solidarity, efficiency and quality of care.

§ 5.1.2 Study objective and contribution

This study explores the predictive power of interactions in the Dutch RE-model of 2014. This RE-model include four morbidity-based risk adjusters, which are a risk adjuster for prior use of specific 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 three preceding years (MHC-groups). Due to the sophistication of a 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 them are relevant from a statistical point of view. We use regression trees to automatically identify interactions, preventing exhaustive hand searches. The regression trees identify interactions that *statistically significantly* explain variation in observed expenses that is not already explained by the risk adjusters in the RE-model. The identified interactions are then used as additional risk adjusters in the RE-model. By comparing the predictive performance of the extended RE-models with the predictive performance of the Dutch RE-model of 2014 we conclude to what extent interactions improve the predictive performance of this RE-model; 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 healthcare expenses. For the purpose of this study, we do not use regression trees to predict costs per se but used them to identify possible interactions in the residuals of the RE-model. In a second step, the identified interactions are used as additional risk adjusters in the RE-model, given the Ordinary Least Squares (OLS)-model as used in practice. Buchner and colleagues (2014) were the first who used 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), and we control for overfitting (see § 5.2). Second, instead of estimating a single tree, several trees are estimated to test the robustness with respect to the identification of interactions. Third, due to availability of external information; i.e. information that is not explicitly used in developing the RE-model, we were able to assess model's predictive performance on selected non-random groups in order to measure financial incentives for risk selection with and without the identified interactions included in the RE-model.

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

Section 5.2 describes the data and methodology. The results of the empirical analysis are presented in Section 5.3. Section 5.4 concludes and discusses the results.

§ 5.2 DATA AND METHODS

§ 5.2.1 Administrative data & health survey data

Dutch administrative data of 2011 was used to develop the regression trees and estimate the RE-models. Information on total healthcare 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 healthcare expenses included all costs related to the Dutch basic benefit package, except mental healthcare services[1]. Total expenses were annualized and weighted by the fraction of the year the individual was enrolled; e.g. an individual who was enrolled for 6 months and had € 500 expenses was given a weight of 0.5 and € 1,000 annual expenses.

A Dutch health survey, "Gecon", was used to assess models' predictive performance at selected non-random groups. This survey is conducted each year on a representative sample of the Dutch population by "Statistics Netherlands"[2]. The survey is targeted on private households. Individuals in mental healthcare 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)[3]. Information on self-reported health status and healthcare utilization were used to select the non-random groups.

§ 5.2.2 How do regression trees work?

Regression trees, developed by Breiman and colleagues in 1984, are non-parametric techniques and belong to the family of "Classification and Regression Trees". 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. Below we will briefly explain how regression trees work. For a thorough discussion of the technical details of regression trees and classification procedures see Breiman et al., 1984; Hastie et al., 2009; and Strobl et al., 2009.

Developing a regression tree starts with *growing* the total tree, which basically means that the data are recursively partitioned into subgroups (Hastie et al., 2009; Sarma, 2007; Berk, 2006). 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

[1] 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 healthcare services. Expenses related to mental healthcare services have been excluded, because in the Netherlands a separate RE-model with specific mental healthcare risk adjusters is applied.

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

[3] 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' healthcare expenses, using information known prior to the estimation year (= 2011).

results into two or more subgroups that are as homogeneous as possible with respect to the target variable (within-variation) and that are maximally differentiated from the other subgroups in terms of the target variable (between-variation). Whether the data at each step is split in two subgroups (i.e. a binary tree) or in more subgroups (i.e. a multi-way tree) depends on a user-defined model parameter. At all next steps of the splitting process, the tree searches the best split, given ancestor split(s). This process continues until pre-specified stopping rules are met, such as the statistical significance of the *F*-statistic[4] or a minimum number of observations per subgroup. In Breiman's 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 can occur, meaning that the leafs describe noise rather than the underlying relationship in the data (Hastie et al., 2009; Berk, 2006). To prevent overfitting, the trees can be pruned. *Pruning* is a process that sequentially removes leafs from the bottom of the total tree and selects the subtree with the highest accuracy (Sarma, 2007). Accuracy is measured in terms of the weighted average of the average squared errors of all leafs of the subtree (Sarma, 2007).

§ 5.2.3 Data preparations for regression trees

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' predictive performance. To prevent overfitting of the trees, we used one sample for growing the trees and another sample for pruning the trees[5]. The leafs of the pruned trees were used for defining the interactions. A third sample was needed to prevent overfitting 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 make maximum use of 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

[4] 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 might 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.

[5] 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. This test shows that the removed leafs do not have high predictive power: the R-squared (R^2) and Cumming's Prediction Measure (CPM) of the RE-model with the interactions of a pruned tree versus those of a total tree dropped 0.01 percentage points and the 'Mean Absolute Prediction Error' (MAPE) remained the same. The removed leafs may describe well patterns in the sample that is used for growing the tree but may not generalize well in another sample.

contained ~25% of the total observations. Consistent with other studies – e.g. Robinson 2008, Hastie et al., 2009 – more data was reserved for growing the tree than for pruning or evaluation, because this generally results in more stable estimates of the tree (Sarma, 2007).

In the total administrative dataset, mean observed expenses were € 1,785 and mean residual expenses were € 0[6] (Table 5.1). 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 a 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 5.1, our split sampling procedure did not yield bias in the representativeness of the administrative samples (see Appendix 5.1 for detailed statistics).

Table 5.1: 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 [a]

	Administrative dataset				Health survey sample
	Total dataset	**Sample 1**	**Sample 2**	**Sample 3**	
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 expenses in €'s (std.) [b, c]	1,785 (5,978)	1,783 (5,944)	1,785 (6,131)	1,787 (5,885)	1,766 (5,364)
Median total observed expenses in €'s	445	445	445	446	444
Mean age in years (std.) [c]	40.078 (22.924)	40.050 (22.934)	40.069 (22.926)	40.142 (22.902)	39.687 (22.986)
Proportion male	0.493	0.493	0.493	0.493	0.486
Proportion classified in a PCG [d]	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 [e, f]	0.087	0.086	0.086	0.087	0.087
Proportion classified in a MHC-group [g]	0.058	0.058	0.058	0.058	0.059
Proportion classified in a DME-group [h]	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

Footnotes Table 5.1:

a. All statistics are weighted for the enrolment period of individuals.
b. Observed expenses are annualized and weighted for the enrolment period in 2011. All expenses are rounded to the nearest €.
c. To calculate the standard deviation, the sum of the weights minus one is used as the variance divisor.
d. PCG: Pharmaceutical Cost Group.
e. DCG: Diagnostic Cost Group.
f. Individuals can be classified in only one DCG, the one with the highest follow-up costs.
g. MHC-group: Multiple-year High Cost-group.
h. DME-group: Durable Medical Equipment-group.

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

To develop the trees, all individuals who were *not* enrolled the full year were excluded: ~3% in sample 1 and sample 2 (Table 5.2, Appendix 5.2 provides detailed descriptive statistics). 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 new borns, 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.

Table 5.2: Descriptive statistics of the total administrative dataset and sample 1 and sample 2 of the administrative dataset from the Dutch population of insured in 2011, *after* exclusion of the individuals who were *not* enrolled the full year, which were *only* used for developing the regression trees to identify interactions [a]

	Total dataset	**Sample 1**	**Sample 2**
N(individuals) [b]	16,166,845	8,065,188	4,114,567
N(insured-years) [b]	16,166,845	8,065,188	4,114,567
Number of excluded individuals	522,116	262,392	133,079
Mean total observed healthcare expenses in €'s (std.) [c]	1,693 (4,999)	1,691 (5,027)	1,692 (4,962)
Median total observed healthcare expenses in €'s	442	442	443
Mean age in years (std.)	40.226 (22.734)	40.202 (22.743)	40.217 (22.736)
Proportion male	0.493	0.493	0.493
Proportion classified in a PCG [d]	0.173	0.173	0.173
Proportion classified in multiple PCGs	0.034	0.034	0.034
Proportion classified in a DCG [e, f]	0.086	0.086	0.086
Proportion classified in a MHC [g]	0.057	0.057	0.057
Proportion classified in a DME [h]	0.008	0.008	0.008
Proportion classified in a PCG, DCG, MHC, and/or DME	0.221	0.221	0.221

Footnotes Table 5.2:

a. Statistics for sample 3 of the administrative dataset are not reported in this table, since this sample is not used for developing the regression trees. Statistics of the total administrative dataset are reported to indicate the representativeness of the sample 1 and sample 2, after exclusion of the individuals who were enrolled for a part of the year.
b. The number of individuals is equivalent to the number of insured-years, since all individuals who were enrolled for a part of the year were excluded.
c. Observed expenses are annualized and weighted for the enrolment period in 2011. All expenses are rounded to the nearest €.
d. PCG: Pharmaceutical Cost Group.
e. DCG: Diagnostic Cost Group.
f. Individuals can be classified in only one DCG, the one with the highest follow-up costs.
g. MHC-group: Multiple-year High Cost-group.
h. DME-group: Durable Medical Equipment-group.

§ 5.2.4 Regression tree model specifications

The target variable in the trees was residual expenses[7,8], defined as observed expenses minus predicted expenses by the Dutch RE-model of 2014[9]. Using this definition, a positive value implies an under-prediction of expenses and a negative value an over-prediction of expenses. With our definition of the target variable, the tree finds interactions that explain variation in observed expenses that is not already explained by the risk adjusters in the 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 taken into account[10].

The key variables in the trees were the risk adjusters in the Dutch RE-model of 2014, which were a categorical variable for age interacted with gender (40 classes), source of income interacted with age (18 classes), region (10 classes), socioeconomic status interacted with age (12 classes), DCGs (16 classes), MHC-groups (7 classes), DME-groups (5 classes), and a binary variable for each of the 24 PCGs. Table A.1 provides the definition of these risk adjusters (see page 255). 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, since binary variables have less flexibility in defining splits than categorical variables; i.e. 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.

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 of the sample (Gail et al., 2009; Hastie et al., 2009; Last et al., 2002), we grew *Tree 2* on sample 2 and pruned it on a random half of

[7] Residual healthcare 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; e.g. log-Euros.

[8] 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.

[9] This model was estimated on the total administrative dataset. The dependent variable was observed expenses and the independent variables were the eight risk adjusters (see Table A.1, page 255). The model used a constant and a weight for the enrollment period.

[10] An additional test, where we estimated a tree with observed expenses as the target variable, ceteris paribus, showed that far more interactions were identified than a tree with residual expenses as the target variable: 327 versus 128 interactions. However, these 328 interactions have about a similar predictive power: the R^2 of an RE-model with these 328 interactions is 0.06 percentage points higher than an RE-model with 128 interactions, the CPM is 0.13 percentage points higher, and the MAPE is € 2 lower.

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 of the sample 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 3 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 to be used in practice.

Next to the aforementioned model specifications, it was required to specify some additional parameters (see Appendix 5.3). 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 user-defined 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, since 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 of the trees a node can be divided in 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 multi-way splits can be achieved by several binary splits (Hastie et al., 2009). We allowed risk classes to be used multiple times across ancestor splits: e.g. 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 individuals can be used again to further split this group 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 5.4).

§ 5.2.5 Risk equalization models

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

§ 5.2.6 Comparative model evaluation

All RE-models were evaluated for all individuals in sample 3 and for selected non-random groups. At the sample level, the 'adjusted R-squared' (R^2), 'Cumming's Prediction Measure' (CPM), and 'Mean Absolute Prediction Error' (MAPE) were calculated for each RE-model. The R^2 was calculated as: one minus the ratio of the variance of the error to the variance of observed expenses, adjusted for the number of risk classes in the model. The CPM was calculated as: one minus the ratio of the mean absolute difference between predicted expenses and observed expenses to the mean of the absolute difference between individual observed expenses and average observed expenses. The MAPE was calculated as: the mean of the absolute difference between predicted expenses and observed expenses. These measures-of-fit examine how well the models on average predict expenses for the total sample. For a thorough discussion of these measures-of-fit, see van Veen et al., 2015a (see Chapter 2).

The 'Mean Prediction Error' (MPE) was calculated for 46 selected non-random groups, using an approach similar to other studies (van Kleef et al., 2014; Stam et al., 2010b). The MPE was calculated as: the mean of predicted expenses minus observed expenses. 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 select the groups. Most groups were defined by 'yes/no'-questions. Table A.2 describes the definition of the evaluation-groups (see page 257). The groups are non-random, since they comprised an over-representation of high-risk individuals; e.g. 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 calibrated in such a way that the overall MPE for each model was zero: the MPE were 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 5.1). 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.

§ 5.3 RESULTS

§ 5.3.1 Robustness of the identified interaction terms

Given the defined tree specifications, Tree 1 to Tree 5 identify 128, 105, 91, 145, and 7 interactions, respectively. Many higher-order interactions with different levels are identified. 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 eleventh-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 Tree 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. These findings show that the trees are not robust in terms of the identification of interactions.

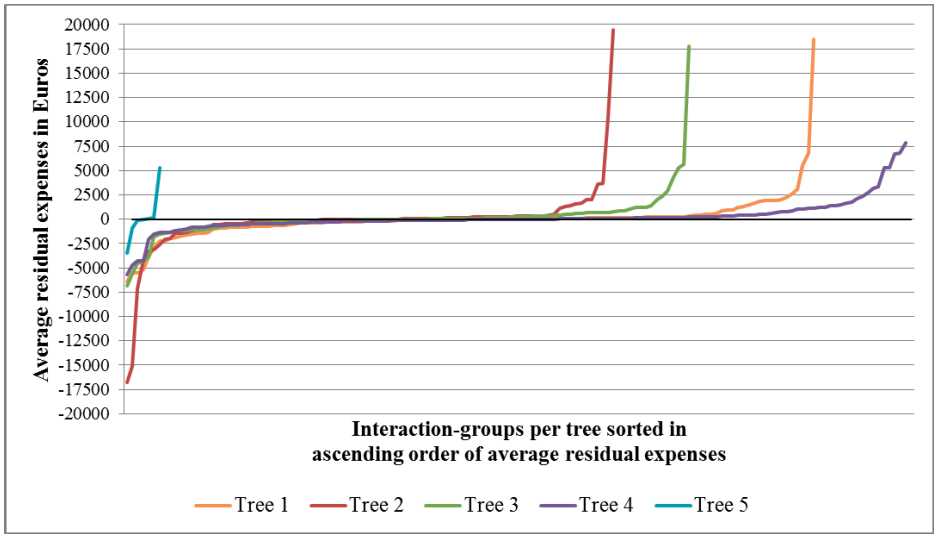
§ 5.3.2 Interaction-groups

Figure 5.1 presents the average residual expenses for the interaction-groups for all trees calculated on sample 3. This figure shows that the trees identified groups with substantial residuals, implying that the RE-model does not adequately predict expenses for these groups. Consequently, the residuals for these groups will be close to zero when interactions are included in the model, because of estimating the RE-model by OLS.

§ 5.3.3 Models' predictive performance at the sample level

Table 5.3 shows that interactions can improve model's predictive performance at the sample level. The R^2 of RE-model 0 is 25.56% and the CPM is 24.98%. Including interactions in this model increases the R^2 by 0.08 to 1.78 percentage points and the CPM increases 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 identify high-risk individuals because the R^2 is more sensitive to large prediction errors than the CPM. 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

Figure 5.1: Average residual expenses for the interaction-groups that were identified by each of the five estimated regression trees, with the interaction-groups per tree being sorted by their average residual expenses [a,b]



Footnotes Figure 5.1:

a. The average residual expenses per interaction-group cannot be compared across the trees, since they can be different interaction-groups with the same average residual expenses. The above figure illustrates the distinctive power of each trees in terms of indentifying groups that are under-predicted and over-predicted by the Dutch RE-model of 2014.

b. For the predictive performance of the RE-model in terms of the R^2, CPM, and MAPE as presented in Table 5.3, all interaction-groups are relevant (and not only the interaction-groups with the highest and lowest average residual expenses).

Table 5.3: 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) [a]

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

Footnotes Table 5.3:

a. 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).

b. 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.

c. CPM = Cumming's 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.

d. 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 €.

reduces when interactions are included in the model. Of all extended RE-models, RE-model 4 performs best in terms of R^2, CPM, and MAPE at the sample level and therefore, we examine to what extent this model predicts expenses for the selected non-random groups.

§ 5.3.4 Models' predictive performance for selected non-random groups

For all selected non-random groups average observed expenses are (far) above average observed expenses in the survey sample, indicating that the groups contain a high proportion of relatively high-risk individuals. A negative MPE on a group in Table 5.4 implies a positive MPE on the complementary group and vice versa. Note that individuals can be classified to multiple groups.

Table 5.4 shows that there are 18 groups for which the MPE is statistically significantly different from zero for RE-model 0; e.g. this model under-predict expenses for persons who have limitations in hearing, a low score on physical health scales, or a poor general health status (Table 5.4). We examined whether including interactions in this model may reduce the MPE at these groups. Table 5.4 shows that RE-model 4 reduces the MPE for individuals who contacted a home nurse or who used durable medical equipment in the past year. However, the MPE for the other groups are still statistically significantly different from zero,

Table 5.4: The Mean Prediction Error (MPE) 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 (N = 16,141) [a, b, c]

Groups (based on health survey data from 2010)	Size, in %	Mean observed expenses in 2011, in €'s	MPE in 2011, in €'s	
			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
Use of durable medical equipment	7.2	5,094	-527*	-455
General health status (all respondents)				
General health status is poor	19.0	4,279	-380***	-391***
At least one long-term disease	31.6	3,480	-334***	-321***
Functional disabilities (age ≥ 12 years)				
OECD limitations in seeing	6.1	3,883	-712**	-694**
OECD limitations in hearing	2.9	5,133	-1,207**	-1,232**
Scores on SF-12 (age ≥ 12 years)				
A low score on physical health scales	19.0	4,476	-671***	-674***
The lowest score on physical health scales	9.5	5,707	-835**	-859**
Self-reported disease in the past year (age ≥ 12 years)				
Serious / persistent back problems or pain	10.6	3,488	-368**	-358*
Serious bowel disorders, longer than 3 months	3.4	4,412	-745*	-705*
Co-morbidity (age ≥ 12 years)				
Three or more self-reported diseases or (chronic) disorder	17.5	4,212	-336**	-361**

Table 5.4: (continued)

Groups (based on health survey data from 2010)	Size, in %	Mean observed expenses in 2011, in €'s	MPE in 2011, in €'s	
			Model 0	Model 4
Health care utilization (all respondents)				
Hospitalization in the past year	6.6	5,773	-576**	-596**
Prescribed drugs use in the past 14 days	35.7	3,133	-188***	-171**
Contact general practitioner in the past year	72.0	2,068	-81*	-84**
Contact medical specialist in the past year	37.9	3,114	-326***	-326***
Contact physiotherapist in the past year	21.8	2,934	-327***	-328***
Hearing-aid	3.4	4,760	-613*	-614*
Limitations in daily activities (ADL) (age ≥ 55 years)				
At least one bad score on ADL scales	3.6	7,227	-640*	-638*

Footnotes Table 5.4:

a. MPE = Mean Prediction Error, calculated as mean of (predicted expenses minus observed expenses).

b. ***: Statistically significantly different from zero with *P*-value ≤ 0.01; **: Statistically significantly different from zero with *P*-value ≤ 0.05; *: Statistically significantly different from zero with *P*-value ≤ 0.10 (based on an one-sample two-sided *T*-test).

c. Predicted expenses in the survey sample were calibrated in such a way that average MPE (=predicted expenses minus observed expenses) on the total survey sample are 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.

implying that RE-model 4 still does not adequately predict expenses for several non-random groups. The other extended RE-models provide similar results[11]. For 28 groups, 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 5.5).

§ 5.4 CONCLUSIONS AND DISCUSSION

This study explored the predictive power of interaction terms between the risk classes in the Dutch risk equalization (RE) model of 2014 using regression trees. Several regression trees were developed to investigate the robustness of the tree in terms of the identified interactions. The identified interactions were used as additional risk adjusters in the Dutch RE-model of 2014. The predictive performance of these extended RE-models was compared to the predictive performance of the Dutch RE-model of 2014. This study has three important results.

First, we found evidence that interactions can improve models' predictive performance. 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

[11] For ease of interpretability, we did not present the results for all RE-models on all selected non-random groups.

MAPE of € 1,569 decreases by € 0 to € 9, depending on the specification of the regression tree.

Second, our analysis shows that 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. This finding is consistent with the literature (Gail et al., 2009; Hastie et al., 2009; Strobl et al., 2009). Consequently, we cannot draw conclusions about which interactions should be used in practice. To decide on which interactions to be used, other criteria may 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 expert, e.g. interactions reflecting co-morbidity.

Third, we show that interactions can reduce financial incentives for risk selection but cannot eliminate them. An RE-model with interactions still does not adequately predict expenses for some selected non-random groups. Despite of this, it is noteworthy to mention that the residuals will be (close to) zero for all groups that are explicitly distinguished in the RE-model, given the use of Ordinary Least Squares. Consequently, including interactions in the RE-model lead to adequate predictions of average expenses for all groups in the RE-model and so, financial incentives for risk selection on these specific groups are removed completely.

Appendices
Chapter 5

APPENDIX 5.1: DETAILED DESCRIPTIVE STATISTICS (I)

Table A.5.1: Descriptive statistics of the administrative dataset from the Dutch population of insured in 2011 (N = ~16.7 million), the three samples of this dataset used for the statistical analysis, and the health survey sample used for model evaluation

	Total administrative dataset	Samples of administrative dataset			Health survey sample
		Sample 1	Sample 2	Sample 3	
N(individuals)	16,688,961	8,327,580	4,247,646	4,113,735	16,141
N(insured-years) [a]	16,438,958	8,201,696	4,184,047	4,053,215	16,067
Healthcare expenses					
Mean total observed expenses, in €'s [b,c]	1,785 (5,978)	1,783 (5,944)	1,785 (6,131)	1,787 (5,885)	1,766 (5,364)
Median total observed expenses, in €'s	445	445	445	446	444
% with predicted expenses < mean total observed expenses	73.24%	73.23%	73.23%	73.26%	74.64%
% with predicted expenses ≥ mean total observed expenses	26.76%	26.77%	26.77%	26.74%	25.36%
Age/gender					
Men 0-24 years	15.07%	15.09%	15.07%	15.05%	15.38%
Men 25-44 years	13.00%	13.01%	13.01%	12.98%	11.37%
Men 45-64 years	14.15%	14.13%	14.17%	14.17%	14.50%
Men 65-74 years	4.36%	4.35%	4.35%	4.38%	4.70%
Men ≥75 years	2.76%	2.76%	2.75%	2.78%	2.66%
Women 0-24 years	14.48%	14.50%	14.48%	14.43%	15.70%
Women 25-44 years	13.08%	13.08%	13.07%	13.08%	12.56%
Women 45-64 years	14.11%	14.09%	14.09%	14.15%	14.86%
Women 65-74 years	4.60%	4.59%	4.62%	4.59%	4.65%
Women ≥75 years	4.40%	4.39%	4.39%	4.41%	3.62%
Region					
Cluster 1-5	49.57%	49.60%	49.60%	49.49%	47.23%
Cluster 6-10	50.43%	50.40%	50.40%	50.51%	52.77%
Source of income					
Individuals <18 years or >64 years	37.23%	37.26%	37.24%	37.14%	39.35%
Disability benefit	4.96%	4.96%	4.98%	4.95%	4.10%
Social security benefit	1.97%	1.97%	1.98%	1.94%	1.27%
Student	3.28%	3.27%	3.27%	3.31%	3.36%
Self-employed	4.05%	4.05%	4.06%	4.05%	3.50%
Others	48.51%	48.48%	48.47%	48.61%	48.24%
Socio-economic status					
Living on a home address with ≥15 persons	1.21%	1.22%	1.21%	1. 19%	0.34%

Table A.5.1: (continued)

	Total administrative dataset	Samples of administrative dataset			Health survey sample
		Sample 1	Sample 2	Sample 3	
Lowest income-class (deciles 1-3)	29.63%	29.65%	29.64%	29.65%	26.91%
Middle income-class (deciles 4-7)	39.52%	39.53%	39.51%	39.53%	40.82%
Highest income-class (deciles 8-10)	29.64%	29.60%	29.64%	29.71%	31.94%
Durable-medical equipment					
No equipment	99.19%	99.19%	99.20%	99.18%	99.09%
Insulin pump	0.11%	0.11%	0.11%	0.11%	0.15%
Catheter	0.39%	0.39%	0.39%	0.39%	0.43%
Colostomy	0.29%	0.29%	0.29%	0.30%	0.31%
Trachea-colostomy	0.02%	0.02%	0.02%	0.02%	0.02%
Multiple year high-costs					
No multiple year high costs	94.23%	94.23%	94.24%	94.21%	94.12%
2-years top 10%	1.00%	1.00%	1.00%	1.00%	1.10%
3-years top 15%	2.31%	2.31%	2.30%	2.31%	2.31%
3-years top 10%	1.06%	1.06%	1.06%	1.07%	1.00%
3-years top 7%	0.80%	0.79%	0.80%	0.80%	0.80%
3-years top 4%	0.46%	0.46%	0.45%	0.46%	0.49%
3-years top 1.5%	0.15%	0.15%	0.15%	0.16%	0.17%
% classified in one or more PCGs	17.30%	17.29%	17.30%	17.33%	17.22%
% classified in multiple PCGs	3.47%	3.47%	3.46%	3.48%	3.46%
% classified in a DCG	8.65%	8.64%	8.64%	8.69%	8.72%
Combinations of risk classes					
% classified into a PCG, DCG, DME-group, and/or MHC-group	22.05%	22.03%	22.04%	22.09%	22.02%
% *not* classified into a PCG, DCG, DME-group, and MHC-group	77.95%	77.97%	77.96%	77.91%	77.98%

Footnotes Table A.5.1:

a. This is the sum of the weights for the fraction of the year the individual was enrolled. The number of insured-years is lower than the number of individuals, because not all individuals were enrolled the full year.

b. Expenses are annualized and weighted for the enrolment period. All expenses are rounded to the nearest €.

c. Standard deviation is presented in parentheses.

APPENDIX 5.2: DETAILED DESCRIPTIVE STATISTICS (II)

Table A.5.2: Descriptive statistics of the samples of the administrative dataset from the Dutch population of insured in 2011, *after* exclusion of the individuals who were *not* enrolled the full year, used for developing the regression trees [a]

	Total administrative dataset	Samples of administrative dataset in 2011	
		Sample 1	Sample 2
N(individuals)	16,166,845	8,065,188	4,114,567
N(insured-years) [b]	16,166,845	8,065,188	4,114,567
Number of excluded individuals	522,116	262,392	133,079
Healthcare expenses			
Mean total observed expenses, in €'s [c,d]	1,693 (4,999)	1,691 (5,027)	1,692 (4,962)
Median total observed expenses, in €'s	442	442	443
% with predicted expenses < mean total observed expenses	72.55%	72.52%	72.52%
% with predicted expenses ≥ mean total observed expenses	27.45%	27.48%	27.48%
Age/gender			
Men 0-24 years	14.90%	14.92%	14.90%
Men 25-44 years	13.02%	13.04%	13.03%
Men 45-64 years	14.29%	14.27%	14.31%
Men 65-74 years	4.38%	4.38%	4.38%
Men ≥75 years	2.69%	2.69%	2.68%
Women 0-24 years	14.32%	14.34%	14.33%
Women 25-44 years	13.15%	13.16%	13.15%
Women 45-64 years	14.28%	14.26%	14.26%
Women 65-74 years	4.64%	4.64%	4.66%
Women ≥75 years	4.32%	4.31%	4.31%
Region			
Cluster 1-5	49.50%	49.52%	49.53%
Cluster 6-10	50.50%	50.48%	50.47%
Source of income			
Individuals <18 years or >64 years	36.81%	36.84%	36.83%
Disability benefit	5.01%	5.01%	5.03%
Social security benefit	1.96%	1.97%	1.97%
Student	3.31%	3.30%	3.29%
Self-employed	4.09%	4.08%	4.10%
Others	48.82%	48.79%	48.78%
Socio-economic status			
Living on a home address with ≥15 persons	1.11%	1.12%	1.11%
Lowest income-class (deciles 1-3)	29.54%	29.57%	29.56%
Middle income-class (deciles 4-7)	39.60%	39.60%	39.59%
Highest income-class (deciles 8-10)	29.75%	29.71%	29.75%

Table A.5.2: (continued)

	Total administrative dataset	Samples of administrative dataset in 2011	
		Sample 1	Sample 2
Durable-medical equipment			
No equipment	99.21%	99.21%	99.22%
Insulin pump	0.11%	0.11%	0.11%
Catheter	0.38%	0.37%	0.37%
Colostomy	0.28%	0.28%	0.28%
Trachea-colostomy	0.02%	0.02%	0.02%
Multiple year high-costs			
No multiple year high costs	94.30%	94.30%	94.31%
2-years top 10%	0.99%	0.99%	0.99%
3-years top 15%	2.30%	2.30%	2.29%
3-years top 10%	1.05%	1.05%	1.05%
3-years top 7%	0.78%	0.78%	0.78%
3-years top 4%	0.44%	0.44%	0.44%
3-years top 1.5%	0.15%	0.14%	0.15%
% classified in one or more PCGs	17.34%	17.33%	17.33%
% classified in multiple PCGs	3.43%	3.43%	3.42%
% classified in a DCG	8.61%	8.60%	8.60%
Combinations of risk classes			
% classified into a PCG, DCG, DME-group, and/or MHC-group	22.09%	22.08%	22.09%
% *not* classified into a PCG, DCG, DME-group, and MHC-group	77.91%	77.92%	77.91%

Footnotes Table A.5.2:

a. The results for sample 3 are not reported in this table, because this sample is not used for developing the regression trees.
b. The number of insured-years is equivalent to the number of individuals, since all individuals who were enrolled for a part of the year were excluded.
c. Expenses are annualized and weighted for the enrolment period. All expenses are rounded to the nearest €.
d. Standard deviation is presented in parentheses.

APPENDIX 5.3: REGRESSION TREE PARAMETERS

This appendix describes the parameters that are required specifying within SAS® Enterprise Miner 12.1 to estimate the regression trees.

All specified tree parameters:

Train
- Use frozen trees: No (default)
- Use multiple targets: No (default)
- Precision: 4 (default)

Splitting rules
- Interval criterion: Prob*F* (default) Note: this study uses an interval variable
- Nominal criterion: Prob*Chisq* (default)
- Ordinal criterion: Entropy (default)
- Significance level: 0.05
- Missing values: Use in search (default) Note: there are no missing values in our data
- Use input once: No (default) Note: key variables can be used multiple times
- Maximum branch: 2 (default) Note: this parameter defines a binary tree
- Maximum depth: 50
- Minimum categorical size: 5 (default)
- Split precision: 4 (default)

Node
- Leaf size: 862 for Trees 1, 4, and 5; 432 for Tree 2; and 415 for Tree 3
- Number of rules: 5 (default)
- Number of surrogate rules: 0 (default)
- Split size: 863 for Trees 1, 4, and 5, 433 for Tree 2; and 416 for Tree 3

Split search
- Use decisions: No (default)
- Use priors: No (default)
- Exhaustive: 5000 (default)
- Node sample: 8,065,188 for Trees 1, 4, and 5; 4,114,567 for Tree 2; and 4,033,033 for Tree 3

Appendix 5.3: (continued)

Subtree
- Method: Assessment (default) Note: this parameter defines a pruned tree
- Assessment measure: average squared error

Cross validation
- Perform cross validation: No (default)

Observation based importance
- Observation based importance: No (default

P-value adjustment
- Bonferroni adjustment: Yes (default)
- Time of Kass adjustment (i.e. Bonferroni adjustment): before split (default)
- Inputs: No (default)
- Split adjustment (i.e. depth adjustment): Yes (default)

Output variables
- Leaf variable: Yes (default)
- Performance: disk (default)

Score
- Variable selection: Yes (default)
- Leaf role: segment (default)

Arguments for specifying other values the default:

Significance level for testing the *F*-statistics
- This parameter is by default 0.2. The purpose of our study is to find interaction terms that significantly reduce risk heterogeneity in residual expenses and so, significantly contribute to model's predictive performance. Since we have a large dataset and many relevant variables, we used a more stringent *P*-value than the default value for testing the statistical significance of splits.

Maximum depth
- This parameter is by default 6. Sine we use a large dataset and many relevant variables, it was expected that higher-than-sixth-order interactions may arise. To avoid the situation

Appendix 5.3: (continued)

that the tree stops with splitting nodes, while there may be useful higher-than-sixth-order terms, we specified a maximum depth of 50. This value would be large enough to not restrict the depth of the tree. Consequently, the tree stops when the split is not statistically significant or the size of the node would be smaller than the specified minimum size.

Leaf size

- This parameter is by default 5. For the purpose of our study, stability of groups is important. Therefore, we specified a minimum leaf size that is equal to the number of individuals in the smallest risk class in the Dutch RE-model of 2014, which is 862 individuals for Trees 1, 4, and 5 (sample 1, after exclusion of individuals who were not enrolled the full year), 432 individuals for Tree 2 (sample 2, after exclusion of individuals who were not enrolled the full year), and 415 individuals for Tree 3 (random half of sample 1, after exclusion of individuals who were not enrolled the full year).

Split size

- This parameter is by default set to missing. This means that a node containing a missing value is enough to split the node. In our sample, we do not have missing values for the variables. Therefore, we set this parameter to 863 for Trees 1, 4, and 5, implying the procedure considers a node for splitting if this node has 863 or more observations. This parameters works together with the leaf size parameter. Note that a node with 863 observations will not be split further, since then the node size after the split would be smaller than the required minimum. The split size for Tree 2 was set to 433 and for Tree 3 to 416.

Node sample

- This parameter is by default 20,000, implying that if the number of observations in a node is larger than 20,000, the split search for that node is based on a random sample of this size. We have a large number of observations in the dataset and so, we have risk classes with a much larger size of 20,000. We did not want to have random splits and therefore, we set this value equal to the total number of observations in the sample used for growing the tree.

Appendix 5.3: (continued)

Assessment measure

- This parameter is by default set to decision, which selects a tree that has the largest average profit and smallest average loss, if a profit-loss matrix is defined. If no profit-loss matrix is defined and the target variable is interval, then the procedure uses average squared error as the assessment measure. We do not use a profit-loss matrix and the target variable is interval. So, we set the assessment measure directly to average squared error.

APPENDIX 5.4: RESULTS OF THE SENSITIVITY ANALYSIS

Table A.5.3 shows that the alternative specification of the significance level and minimum leaf size result into a different set of identified interactions, except for the use of corrected *P*-values. However, the predictive performance of alternative RE-model 1 and 2 is similar to the predictive performance of RE-model 1. Alternative RE-model 3 has a higher predictive performance than RE-model 1, however, this alternative RE-model uses the interactions that are identified by a tree with a minimum leaf size of 1. For reasons of stability, a risk class with 1 observation will not be used in an RE-model.

Table A.5.3: Results of regression trees with an alternative specification for each of the user-defined parameters, ceteris paribus to Tree 1, on *total* sample 3 (*N* = ~4.1 million)

RE-models [a]	Number of identical leafs to Tree 1	Adj.-R^2 [b] (in %)	CPM [c] (in %)	MAPE [d] (in €'s)
Model 1(Model 0 + all 128 interactions by Tree 1)	-	26.20	25.11	1,566
Alternative Model 1 (Model 0 + all 114 interactions by a tree with a 0.01 significance level for testing the *F*-statistic)	103 (= 80%)	26.20	25.11	1,566
Alternative Model 2 (Model 0 + all 128 interactions by a tree with uncorrected *P*-values)	128 (= 100%)	26.20	25.11	1,566
Alternative Model 3 (Model 0 + all 175 interactions by a tree with a minimum leaf size of 1)	28 (= 22%)	27.79	25.51	1,588

Footnotes Table A.5.3:

a. 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).

b. 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.

c. CPM = Cumming's 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.

d. 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 €.

APPENDIX 5.5: THE MEAN PREDICTION ERROR FOR SOME EVALUATION-GROUPS

Table A.5.4: The 'Mean Prediction Error' (MPE) for RE-model 0 and RE-model 4 on groups for which the MPE was not statistically significantly different from zero for RE-model 0, using the health survey sample (*N* = 16,141) [a]

Groups (based on health survey data from 2010)	Size, in %	Mean observed expenses in year 2011, in €'s	MPE in 2011, in €'s	
			Model 0 [b]	Model 4 [c]
General health status (all respondents)				
Obesity	8.7	3,169	-234	-219
Functional disabilities (age ≥ 12 years)				
OECD limitations in moving	7.5	5,819	-536	-501
OECD limitations in talking	0.3	7,534	-1,626	-1,893
Functional disabilities (age ≥ 12 years)				
OECD limitations in eating	4.0	5,118	-624	-616
Scores on SF-12 (age ≥ 12 years)				
The lowest score on mental health scales	9.6	2,732	-278	-266
A low score on mental health scales	19.1	2,461	-114	-149
Presence of disease or disorder (age ≥ 12 years)				
Diabetes mellitus	5.5	5.191	401	301
Stroke, brain infarction (ever)	2.5	5,274	44	-40
Myocardial infarction or other serious heart disease (ever)	3.1	6,425	-600	-615
Some type of cancer (ever)	7.3	4,895	-353	-378
Self-reported disease in the past year (age ≥ 12 years)				
Migraine or serious headaches regularly	15.4	1,925	-12	-13
Hypertension	17.0	3,388	-195	-206
Vascular constriction (in stomach or legs)	2.1	5,178	-363	-336
Asthma, chronic bronchitis, lung emphysema	8.2	3,961	-192	-168
Psoriasis	2.9	2,294	-159	-116
Chronic eczema	4.5	2,510	105	-38
Dizziness with falling down	3.2	3,568	-148	-169
Urine incontinence	5.7	4,310	-460	-468
Arthritis of hips or knees	14.8	3,468	-180	-144
Rheumatoid arthritis	5.5	3,750	-189	-166
Serious /persistent problems of neck or shoulder	10.5	3,107	-253	-230
Serious/persistent problems of hand, wrist or elbow	6.3	3,373	-301	-294
Other long-term disease or disorder	11.5	4,335	-449	-544*
Co-morbidity (age ≥ 12 years)				
Two self-reported diseases or (chronic) disorder	6.5	2,307	6	-15
Health care utilization (all respondents)				

Table A.5.4: (continued)

Groups (based on health survey data from 2010)	Size, in %	Mean observed expenses in year 2011, in €'s	MPE in 2011, in €'s	
			Model 0 [b]	Model 4 [c]
Contact with visiting home nurse (cure) in the past year	0.8	8,865	-1,046	-916
Home help assistance in the past year	1.5	4,247	-549	-312
Glasses or contact lenses	37.1	2,275	66	82
Complete dentures	10.5	4,456	-171	-153

Footnotes Table A.5.4:

a. Predicted expenses in the survey sample were calibrated in such a way that average MPE (=predicted expenses minus observed expenses) on the total survey sample are 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 sample due to the use of an external dataset, which is a subsample of the administrative dataset.

b. RE-model 0 = Dutch RE-model of 2014.

c. RE-model 4 = Dutch RE-model of 2014 + 145 interaction terms identified by Tree 4.

Chapter 6

Residual Expenses from Multiple Prior Years

ABSTRACT

Selective groups of insured that are under-compensated under an RE-model, especially when they persist over time, are vulnerable to risk selection. This chapter explores whether there are individuals with persistent under-compensations over a period of three years under a morbidity-based risk equalization (RE) model. We use a rich cross-sectional time-series administrative dataset covering almost the entire Dutch population to examine the Dutch RE-model of 2013 over a three-year time period. This study makes two important contributions. First, it confirms the existence of individuals who are persistently under-compensated. On average these individuals differ markedly from the total population: they are relatively unhealthy and most of them have multiple long-term diseases. Second, this study shows that extending the RE-model with a risk adjuster or interaction terms defining the persistently under-compensated group can improve model's predictive performance for the full sample and for some selective groups; however, the prediction of expenses for other selective groups may deteriorate. Consequently, financial incentives for risk selection are mitigated but not eliminated. Applying our method to other RE-models than the Dutch RE-model of 2013 may lead to different conclusions about the size and risk characteristics of the persistently under-compensated group and the potential improvement in model's predictive performance. Our method is generally applicable to any RE-model.

§ 6.1 INTRODUCTION

§ 6.1.1 Background

Risk equalization (RE) models are used for calculating risk-adjusted payments to health insurers in several countries worldwide, including Belgium, Germany, Israel, the Netherlands, Switzerland, and the U.S. (van de Ven et al., 2007; Kautter et al., 2014). Via the RE-model, insurers are compensated for predictable variation in individuals' healthcare expenses. In the presence of premium regulation, as is the case in all aforementioned countries, the goal of RE is to mitigate financial incentives for risk selection and thereby to achieve a level playing field for insurers. The extent to which an RE-model mitigates financial incentives for risk selection depends on the model's predictive performance: there are no financial incentives for risk selection for selective groups of interest when average residual expenses (= observed expenses minus RE-predicted expenses) for these groups are (close to) zero[1]. Over the past two decades, the predictive performance of several RE-models that are used in practice has been improved considerably as a result of including morbidity-based risk adjusters relying on inpatient or outpatient diagnostic information, pharmaceutical information, or prior years' expenses (e.g. Adams et al.,2002; Buchner et al., 2013; Ellis & Ash, 1995; Fishman et al., 2003; Hughes et al., 2004; van Kleef & van Vliet, 2012; Kronick et al., 2000; Pope et al., 2000a).

Several studies, however, have shown that even sophisticated morbidity-based RE-models do not predict expenses adequately for several selective groups of interest; for example, individuals who reported functional disabilities or a poor general health status in surveys (e.g. Ash & Byrne-Logan, 1998; Ash et al., 2005; van Kleef et al., 2012a, 2012b, 2013b, 2014; Pope et al., 2000a; van Veen et al., 2015b). Van Kleef and colleagues have found groups for whom the Dutch RE-model on average under-compensates (i.e. positive residual expenses) insurers each year, despite model improvements over time (van Kleef et al., 2012a, 2013b, 2014). Possibly, there are individuals who are persistently under-compensated under a morbidity-based RE-model.

In the literature, it is largely unknown whether there is a group for whom a morbidity-based RE-model persistently under-compensates insurers. Some studies have investigated persistence in *observed* healthcare expenses (e.g. Garber et al., 1998; Monheit, 2003); however, this is not of particular interest because an RE-model with risk adjusters for selective high-cost groups, for example patient groups, adjusts for predictable variation in individuals' observed healthcare expenses. Instead, it may be better to investigate the persistence in *residual* expenses because this indicates how well the RE-model predicts expenses over time.

[1] In practice, average residual expenses for selective groups of interest do not have to equal zero because of transaction costs for engaging in risk selection and uncertainty around the estimates of average residual expenses for groups.

So far, the persistence in residual expenses under a morbidity-based RE-model has been unexplored. Furthermore, if research shows that a morbidity-based RE-model persistently under-compensates a specific group, it is relevant to investigate the risk characteristics of these individuals, because this information can be valuable for improving model's predictive performance and so, mitigating financial incentives for risk selection.

§ 6.1.2 Study objective and contribution

The goal of this study is to explore whether there is a group that is persistently under-compensated under the Dutch RE-model of 2013, and if such a group exists, we aim to explore the costs and risk characteristics of these individuals and to examine to what extent model's predictive performance can be improved by inclusion of an explicit risk adjuster or interaction terms for this group. The Dutch RE-model of 2013 uses a sophisticated set of risk adjusters, including three morbidity-based risk adjusters: prior use of specific drugs in terms of pharmacy-based cost groups (PCGs), prior hospitalization in terms of diagnostic cost groups (DCGs), and multiple-year high costs over the three preceding years (MHC-groups). This RE-model is the most recent Dutch RE-model that can be estimated on the available data from each of the years during our study period of 2008 to 2011. Although our results and conclusions are conditional on this RE-model and the data, the method applied here is generally applicable to any RE-model.

This study uses a rich cross-sectional time-series administrative dataset covering almost the entire Dutch population to identify individuals with persistent under-compensations. This group is based on residual expenses from the years 2008, 2009, and 2010. To identify individuals with persistent under-compensations, it is necessary to use at least two consecutive years in order to leave out people with transitory health problems. We use three instead of two years because when two years are used it can happen that individuals with a short-term health episode or an accident at the end of a calendar year will have high residual expenses in two consecutive years; with three years the chance of these fluctuations are lower. Using three and not more than three years is because not too many individuals may drop out of the analysis because of death.

The main focus of the empirical analysis is on exploring whether a persistently under-compensated group under the Dutch RE-model of 2013 exists. This analysis consists of analyzing patterns in the distributions of residual expenses over a period of four years – three prior years to identify the group plus the estimation-year 2011 to examine whether the pattern that is observed in each of these prior years also occurs in this year – and examining whether there is a group that has a higher probability of being under-compensated than can be expected by pure chance. Our analysis shows that there are indeed individuals with persistent under-compensations under the Dutch RE-model of 2013, whereby alternative group definitions can be used to identify them (see § 6.3.2). Consequently, in the second part of our analysis, we explore the costs and risk characteristics of the groups of persistently

under-compensated individuals. After this explorative analysis, we examine the predictive performance of the Dutch RE-model of 2013 with and without an explicit risk adjuster or interaction terms for the persistently under-compensated groups on the data from 2011 in order to investigate to what extent model's predictive performance can be improved. Model's predictive performance is evaluated on the full sample and on several selective groups of interest that are derived from questions in a health survey. The first evaluation method indicates the overall fit of the model for the total population. The latter evaluation method is used to estimate financial incentives for risk selection for selective groups in the population.

It is of great policy relevance to know whether there is a group that is under-compensated persistently over time under the RE-model that is used in practice and, if such a group exists, to know how large it is in the population and to know specific risk characteristics. Groups for whom an insurer is not adequately compensated for several years are vulnerable to risk selection, which is a potential threat to solidarity, efficiency, and quality of care (Baumgartner & Busato, 2012; Beck et al., 2003; Frank et al., 1998; von Wyl & Beck, 2015; van de Ven & Ellis, 2000). This is the first study providing empirical evidence that such a group exists under the Dutch RE-model of 2013 and attempting to clarify the risk characteristics of these individuals. Further, we examine to what extent the predictive performance of the Dutch RE-model of 2013 can be improved when an explicit risk adjuster or interaction terms defining this specific group is included. Since this risk adjuster and the interaction terms are based on prior years' residual expenses that can be derived from administrative files of (Dutch) insurers, there are no additional costs for collecting new information.

The remainder of this chapter proceeds as follows. The next section describes the data and methods of our empirical analysis. Section 6.3 presents the results. The final section concludes and discusses the results, and provides several health-policy implications.

§ 6.2 DATA AND METHODS

§ 6.2.1 Administrative data and health survey data

Administrative data for the time period 2008 to 2011 for almost the entire Dutch population (N = ~16 million) were used. For each individual and for each year, we had information on total healthcare expenses and risk adjusters, including age, gender, region, source of income, socioeconomic status, pharmacy-based cost groups (PCGs) from prior use of specific drugs, diagnostic cost groups (DCGs) based on prior hospitalization, and multiple year high cost groups (MHC-groups) based on expenses from the three preceding years. Total expenses

per year were the expenses related to the services included in the Dutch basic benefit package in the respective year, except for mental healthcare services[2,3].

Mean total expenses in 2008 to 2011 were € 1,694, € 1,661, € 1,765, and € 1,785, respectively (Table 6.1). In the study population in 2011, average age was 40 years, 17.3% of the individuals are classified into a PCG, 8.7% into a DCG, and 5.8% into a MHC-group. Combining these risk adjusters, 21.9% of the individuals are classified into a DCG, PCG, and/or

Table 6.1: Descriptive statistics of expenses and risk characteristics in the administrative dataset for each year in the time period 2008 to 2011 [a]

	Year 2008	**Year 2009** [g]	**Year 2010** [g]	**Year 2011**
N(records) [b]	15,538,636	15,568,677	16,426,880	16,688,961
N(individuals) [b]	15,468,734	15,513,511	16,364,745	16,688,961
N(insured-years) [b]	15,225,691	15,279,553	16,128,545	16,438,958
Expenses				
Average total observed expenses (std.), in €'s [c]	1,694 (5,651)	1,661 (5,342)	1,765 (5,955)	1,785 (5,978)
Median observed expenses, in €'s	366	398	421	445
Risk characteristics				
Mean age in years (std.)	39.605 (22.790)	39.804 (22.851)	39.847 (22.847)	40.078 (22.924)
Proportion male	0.490	0.491	0.492	0.493
Proportion classified in a PCG [d]	0.161	0.165	0.168	0.173
Proportion classified in multiple PCGs	0.029	0.032	0.033	0.035
Proportion classified in a DCG [e]	0.024	0.025	0.081	0.087
Proportion classified in a MHC-group [f]	0.050	0.073	0.059	0.058
Proportion classified in a PCG, DCG, and/or MHC-group	0.182	0.194	0.213	0.219

Footnotes Table 6.1:

a. All statistics on expenses and risk characteristics were calculated on the total dataset *per year* and were weighted for the enrolment period of individuals in this year. To calculate the standard deviation (= std.), the sum of the weights minus one was used as the variance divisor.
b. The number of individuals was not equal to the number of insured-years because some individuals were not enrolled for the entire year. Further, the number of records was not equal to the number of individuals because some individuals occur multiple times in datasets due to switching from insurer during the year. These duplicate records were merged to one record for each unique individual in the empirical analysis of determining which individuals belong to the persistently undercompensated group or not.
c. Total observed expenses per year were the sum of expenses on services included in the Dutch benefit package per year.
d. PCG: Pharmacy-based Cost Group. Individuals can be classified to multiple PCGs.
e. DCG: Diagnostic Cost Group. Individuals can be classified in only one DCG, the one with the highest follow-up costs.
f. MHC-group: Multiple-year High Cost-group.
g. The total number of individuals from year 2009 to 2010 increased significantly. This was because one insurer did not deliver data in year 2008 and 2009 for calculating the risk-adjusted payments. In year 2010 and 2011, the total dataset covered the total population of insured in the Netherlands.

[2] Mental healthcare expenses were excluded, because in the Netherlands a separate RE-model with different risk adjusters is estimated for these expenses.

[3] Total expenses included the expenses for hospital care, primary care, paramedical care, pharmaceuticals, durable medical equipment, transport in case of illness, dental care, obstetrical care, and maternity care.

MHC-group. These statistics are similar across years, except the percentage of individuals classified to the DCGs in 2008 and 2009 and the MHC-groups in 2009. These differences were caused by some changes in the definition of these risk adjusters during the study period. Ideally, the definition of the risk adjusters is kept constant. We expect that the small changes in the definition of the risk adjusters during our study period do not significantly change our conclusions about the existence of a persistently under-undercompensated group under the Dutch RE-model of 2013.

A Dutch health survey, "Gecon", from 2010 was used to investigate the characteristics of the persistently under-compensated groups that were identified and to evaluate the estimated RE-models with and without an explicit risk adjuster or interaction terms defining the persistently under-compensated groups on selective groups of interest[4]. This survey is conducted each year under a representative sample of the Dutch population by "Statistics Netherlands" in order to collect information on self-reported health status and healthcare utilization[5]. The administrative dataset was merged with the survey respondents at the individual level by using an anonymous identification key (N = 16,141).

§ 6.2.2 Exploring whether a persistently under-compensated group exists

How to identify this group?

As a first step, we calculated residual expenses for each individual in each of the years 2008 to 2010 by estimating the Dutch RE-model of 2013 on the *total administrative dataset per year*. For each year, the dependent variable in this model was annualized total observed expenses and the independent variables were dummy variables for age interacted with gender, source of income interacted with age, region, socioeconomic status interacted with age, PCGs, DCGs, and MHC-groups. A description of these risk adjusters is well-documented elsewhere (van Vliet et al., 2011, 2012; Eijkenaar et al., 2013); see also Table A.1 (see page 255). The model was estimated by Ordinary Least Squares (OLS), with a constant term and a weight for the enrolment period in the respective year; for example, an individual who was enrolled for 6 months and had € 1000 expenses received a weight of 0.5 and € 2000 annualized total expenses[6]. We identify individuals with positive residual expenses in each of the three years. Residual expenses were calculated as observed expenses minus RE-predicted expenses in the respective year.

[4] The reason to use data from 2010 (and not 2011) is to evaluate RE-models, given information that is known a priori the estimation year (= 2011).

[5] "Statistics Netherlands" (="Centraal Bureau voor de Statistiek") is an autonomous agency financed by the Dutch government that collects and analyzes data.

[6] In total, 123, 126, and 128 risk classes were included in the RE-model for the years 2008, 2009, and 2010, respectively. The number of risk classes differs across the years because of small changes in the definition of some risk adjusters.

As a second step, we merged residual expenses from the years 2008 to 2010 with the total administrative dataset from 2011 at the individual level by using an anonymous identification key. For 14.7 million individuals we had prior years' information, which was used to identify individuals with persistent under-compensations; for 1.9 million individuals this information was lacking; for example, for new borns. To identify a group in 2011 that is under-compensated in each of the three previous years, we examined alternative group definitions, such as those individuals with positive residual expenses in the three preceding years or those individuals in the top of the residual distribution in each year of the three preceding years, given that residual expenses were positive (see § 6.3.2). Furthermore, we also estimated the Dutch RE-model of 2013 on the total administrative dataset from 2011 in order to examine whether the distribution of residual expenses that is observed in the years 2008 to 2010 is consistent with the one in 2011 (i.e. the estimation-year for exploring the costs and risk characteristics of the persistently under-compensated groups that were identified and two potential model improvements).

What are their costs and risk characteristics?

For the groups that were identified as "persistently under-compensated in previous years", we examined in detail the costs and risk characteristics and how large this group is in the total population. In the administrative dataset from 2011 we examined the prevalence of the morbidity-based risk adjusters among the individuals with persistent under-compensations and the prevalence of individuals classified to none of these morbidity-based risk adjusters. Further, we examined average observed expenses and average residual expenses for these groups. In order to obtain more detailed information about the risk characteristics of the persistently under-compensated groups, we merged the administrative dataset from 2011 with the survey respondents at the individual level. On this survey sample, we analyzed several descriptive statistics of the persistently under-compensated groups, including the prevalence of individuals with a self-reported long-term disease and average observed expenses and average residual expenses for individuals with and without a long-term disease.

§ 6.2.3 Two potential model improvements

Model estimation

Based on the first part of our analysis, two definitions for the persistently under-compensated group were used for further analysis of two potential model improvements, namely "those individuals with positive residual expenses in each of the three previous years" (*definition type 1*) and "those individuals in the top 50% of residual expenses in each of the three previous years, given those individuals with positive residual expenses in each year" (*defini-*

tion type 2)[7]. Since these definitions may be somewhat arbitrary, we performed a sensitivity analysis in order to test the predictive power of alternative group definitions.

To examine the two potential model improvements, we compared the predictive power of the Dutch RE-model of 2013 to that of the RE-models that were extended with an explicit risk adjuster or interaction terms defining the persistently under-compensated group according to the two groups definitions. This resulted into estimation of the following five models on the total administrative dataset from 2011. First, the Dutch RE-model of 2013 was estimated (*Model 0*), which included the same risk adjusters as previously mentioned (M = number of risk classes = 127)[8]. Second, Model 0 is extended with a dummy variable defining whether an individual belongs to the persistently under-compensated group or not, according to definitions type 1 and type 2, respectively (*Models 1-2*, M = 128). The reference group of this additional risk adjuster included those individuals who were not persistently under-compensated plus those individuals for whom no prior years' information was available. Third, Model 0 is extended with interaction terms between each risk adjuster in Model 0 and a dummy variable for the persistently under-compensated group according to definitions type 1 and 2, respectively, in addition to the inclusion of the main effects (*Models 3-4*, M = 254). The motive for inclusion of interaction terms is that it may lead to more accurate predictions, because average predicted expenses will equal average observed expenses for each risk class in the model for the persistently under-compensated group and the complementary group separately, as a result of estimating the model by OLS. Existing models are estimated on the total population of interest (Buchner et al., 2013; Beck, 2000; Kautter et al., 2014; van Kleef et al., 2013b). Consequently, predicted expenses for each risk class in the model are based on a pooled sample of individuals who are persistently under-compensated and those who are not. If appropriate, non-statistically significant interaction terms can be ignored in practice but here we included all interaction terms. Comparing Models 1 to 4 to Model 0 indicate the predictive power of including a risk adjuster or interaction terms based on residual expenses from the three previous years. Comparing Models 1 and 2 to Models 3 and 4, respectively, indicate the additional value in predictive power when interaction terms are included instead of one extra risk adjuster.

To prevent overfitting, all five RE-models were estimated on a random half of the total administrative dataset from 2011; i.e. the estimation sample. The estimated coefficients were

[7] To calculate the percentiles of residual expenses, the total dataset per year is used, whereby residual expenses were *not* annualized and *not* weighted for the enrollment period. This procedure is analogues to the procedure that is used for calculating the percentiles for defining the MHC-groups in the Netherlands.

[8] Also here the number of risk classes in the Dutch RE-model differ from those of the same RE-model that is estimated on data from each of the three previous years, because of some small changes in the risk adjusters of this model during our study period. Also here we expect that these small changes do not largely influence our conclusions that model's predictive performance can be improved for the full sample and for some selective groups of interest; and that the prediction of costs for some groups may improve, while it may deteriorate for others.

used to predict expenses in the remainder of this administrative dataset; i.e. the validation sample. In order to make efficient use of the survey data, all individuals who were respondents to the survey were first assigned to the validation sample. All remaining individuals in the administrative dataset were randomly assigned to one of the samples. Table 6.2 shows that this split sample approach did not cause bias in the representativeness of the samples with respect to average observed expenses, average age, and the proportion of individuals classified in a PCG, DCG, MHC-group, or a combination of these risk adjusters.

Table 6.2: Descriptive statistics of expenses and risk characteristics in the two samples of the administrative dataset from 2011 and a health survey dataset from 2010 [a, b]

	Administrative data year 2011		Survey sample year 2010
	Estimation-sample	Validation-sample	
N(individuals)	8,327,580	8,361,381	16,141
N(insured-years)	8,201,696	8,237,262	16,067
Expenses			
Average total observed expenses (std.), in €'s [c]	1,783 (5,944)	1,786 (6,011)	1,766 (5,364)
Median observed expenses, in €'s	445	446	444
Risk characteristics			
Mean age in years (std.)	40.050 (22.934)	40.105 (22.914)	39.687 (22.986)
Proportion male	0.493	0.493	0.486
Proportion classified in a PCG [d]	0.173	0.173	0.172
Proportion classified in multiple PCGs	0.035	0.035	0.035
Proportion classified in a DCG [e]	0.086	0.087	0.087
Proportion classified in a MHC-group [f]	0.058	0.058	0.059
Proportion classified in a PCG, DCG, and/or MHC-group	0.219	0.220	0.219

Footnotes Table 6.2:

a. All statistics on expenses and risk characteristics were weighted for the enrolment period of individuals. To calculate the standard deviation (= std.), the sum of the weights minus one was used as the variance divisor.
b. To prevent overfitting of the estimated RE-models, we used a split-sampling procedure. All respondents to the survey were first assigned to the validation-sample in order to make efficient use of this dataset; all remaining individuals were randomly assigned to one of the samples in such a way that they contained approximately 50% of the total dataset in 2011. The estimation-sample was used to estimate the RE-models. The estimated coefficients were used to predict individuals' expenses in the validation-sample. The survey sample was used to evaluate the RE-model on selective groups of interest.
c. Total observed expenses per year were the sum of expenses on services included in the Dutch benefit package per year.
d. PCG: Pharmacy-based Cost Group. Individuals can be classified to multiple PCGs.
e. DCG: Diagnostic Cost Group. Individuals can be classified in only one DCG, the one with the highest follow-up costs.
f. MHC-group: Multiple-year High Cost-group.

Model evaluation

The predictive performance of the estimated RE-models was assessed on the full sample in terms of the R-squared (R^2), Cumming's Prediction Measure (CPM), and the Mean Absolute Prediction Error (MAPE). See van Veen et al. (2015a) how these measures-of-fit are calculated (see also Chapter 2). These measures-of-fit indicate how well the models on

average predict expenses for the full sample. A R^2-value and CPM-value of 1 and a MAPE-value of 0 indicate perfect model fit. It is worth noting that observed expenses were used as the reference point for calculating residual expenses and so, the estimated RE-models cannot, and do not have to, predict expenses perfectly for the full sample because observed expenses include a random component and variation for which the regulator may not desire compensation: e.g. variation in observed expenses due to differences in the efficiency of healthcare delivery (Stam et al. 2010a).

In addition, we examined average residual expenses for several selective groups of interest on the survey sample. A value of zero indicates an adequate prediction of average expenses for this group. The selective groups were derived from questions about self-reported health status and prior healthcare utilization, such as questions like "How do you rate your general health status" or "Do you have one of the following long-term diseases?". In total, 46 groups were defined by using similar definitions as those used in other studies (van Kleef et al., 2012a, 2012b, 2014; Stam, 2007; van Veen et al., 2015b); see also Table A.2 (see page 257). The selected groups contain an over-representation of individuals that are relatively unhealthy because we were interested in the extent to which the RE-models predicted expenses accurately for such groups; e.g. patient groups.

The survey sample is reasonably representative for the Dutch population in terms of the prevalence of a PCG, DCG, and MHC-group, except for individuals in nursing homes or other institutions because these individuals were excluded from the survey (Table 6.2). As a result, average age in the survey sample is lower than in the population. Further, average observed expenses in this sample are lower than those in the population: € 1,766 versus € 1,785 (not statistically different at 5%)[9].

§ 6.3 RESULTS

§ 6.3.1 Persistence in residual expenses

Before analyzing patterns in residual expenses over years, we first analyze the distribution of observed expenses and residual expenses *per year* in order to examine whether there is consistency across years. Table 6.3 clearly shows that the distribution of observed expenses and residual expenses are similar across the years. In total, 21.9% to 23.7% of the population

[9] Residual expenses were calibrated for the differences in average expenses between the survey sample and the population. Individuals' residual expenses *per* RE-model were raised by a factor equaling average RE-predicted expenses in the survey sample divided by average observed expenses in the survey sample. After this calibration, average residual expenses per RE-model are zero in the survey sample, just as is the case in the population. Calibrated residual expenses per RE-model were used to test whether average residual expenses for each selective group are statistically significantly different from zero, based on an one-sample two-sided *T*-test.

has positive residual expenses in a year. For those individuals, average observed expenses are € ~5,100 to € ~5,600 per year. If we focus on a subgroup of all individuals with positive residual expenses, average observed expenses range from more than € ~8,600 for individuals in the top 50% with positive residual expenses in a year to more than € ~64,000 for individuals in the top 1% with positive residual expenses in a year. In addition, total observed expenses in a year are highly concentrated among those individuals with positive residual expenses: they are responsible for ~70% of the total sum of observed expenses in a year and the top 50% of individuals with positive residual expenses (~11% of the population in a year)

Table 6.3: Descriptive statistics of observed expenses and residual expenses per year according to percentiles of the population ranked by residual expenses of the Dutch RE-model of 2013 per year [a, b, c]

Population per year ranked by residual expenses	Year 2008	Year 2009	Year 2010	Year 2011
	Percentage of total population			
Positive residual expenses	21.9	23.2	23.5	23.7
Negative residual expenses	78.1	76.8	76.5	76.3
	Average observed expenses, in €'s [d]			
Positive residual expenses	5,612	5,110	5,374	5,335
Negative residual expenses	615	635	671	693
Positive residual expenses divided into:				
- Top 1%	73,073	64,142	70,299	70,002
- Top 5%	36,385	32,489	34,929	34,426
- Top 10%	25,713	22,934	24,583	23,950
- Top 20%	17,440	15,614	16,617	16,152
- Top 50%	9,597	8,660	9,158	8,994
- Bottom 50%	1,685	1,609	1,650	1,708
	Average residual expenses, in €'s [d]			
Positive residual expenses	3,649	3,218	3,413	3,337
Negative residual expenses	-1,006	-957	-1,035	-1,026
Positive residual expenses divided into:				
- Top 1%	65,332	57,066	62,347	61,599
- Top 5%	30,603	26,936	29,054	28,270
- Top 10%	20,895	18,284	19,645	18,946
- Top 20%	13,602	11,911	12,724	12,265
- Top 50%	6,907	6,072	6,456	6,268
- Bottom 50%	443	405	422	430

Footnotes Table 6.3:

a. Statistics in this table were based on the total administrative dataset *per year*.

b. The rows 'positive residual expenses' and 'negative residual expenses' together form the total dataset per year. The group of individuals with positive residual expenses was split into percentiles: the top 1%, 5%, 10%, 20%, 50% and the bottom 50%. The top 50% and bottom 50% together form the total group of individuals with positive residual expenses in a year.

c. Residual expenses were calculated as: observed expenses minus predicted expenses of the Dutch RE-model of 2013 that was estimated on the total dataset in the respective year.

d. In the calculation of average observed expenses and average residual expenses, we annualized expenses and weighted expenses by the number of insured-years per cell.

are even responsible for ~60% of the total sum of expenses[10]. These statistics indicate that individuals who are under-compensated in a certain year have far above-average observed expenses in that year. Table 6.3 also shows that the Dutch RE-model of 2013 provides large under-compensations: € ~3,200 to € ~3,600 per year, ranging from more than € ~6,000 for individuals in the top 50% with positive residual expenses in a year to more than € ~57,000 for individuals in the top 1% with positive residual expenses in a year.

Given the consistency in the distributions per year, it is of interest to examine to what extent individuals with high residual expenses in a year continue to have high residual expenses in a following year. Table 6.4 displays the persistence in residual expenses. To obtain these statistics, we calculated the actual probabilities of individuals' position in the top 1%, 5%, 10%, 20%, 50% or the bottom 50% of individuals with positive residual expenses in the years 2009, 2010, and 2011, conditional on being in these positions in 2008. We also calculated the expected probability that an individual can be in each of these positions by pure chance. The statistics in Table 6.4 are the ratios of the actual probabilities to the expected probabilities for individuals with the same position a following year as the one in year 2008. These ratios show how many times higher the probability is that an individual is in the same position in both years than can be expected by pure chance. The probabilities conditional on individuals' position in year 2009 and 2010 were also calculated but are not presented in Table 6.4, because these statistics provide a similar pattern (see Appendix 6.1). Furthermore, for simplicity, the ratios for individuals with an increase or decrease in position in a following year are also not presented in Table 6.4; these statistics show that individuals' position may change over time but the probability that individuals with positive residual expenses continue to have positive residual expenses in a following year is still larger than can be expected by pure chance (see Appendix 6.1).

Table 6.4 provides two important findings. First, the probability that individuals stay in the top percentiles of residual expenses declines over the years (ratios in rows), indicating that some individuals in the top percentiles in a year are not in this position anymore in a following year; however, this probability is still larger than can be expected by pure chance (i.e. ratio is greater than one). For example, among those in the top 1% of individuals with positive residual expenses in 2008, the probability that these individuals stay in this position in 2009 is 73.3 times higher than can be expected and in 2010 this is 35.8 times higher than can be expected. Second, the probability that individuals stay in the same top percentiles in both years decreases when an individual is in a lower position in 2008 (ratios in columns). For example, the probability that individuals stay in the top 1% in 2011 is 23.7 times higher than the probability that individuals are in this position by pure chance, compared to 1.8 for individuals in the top 50% of positive residual expenses. Table 6.4 clearly shows that there

[10] These statistics are not presented in Table 6.3.

Table 6.4: Persistence in residual expenses over a four-year time period: ratio of the actual probability that an individual has the same position in a following year to the expected probability that an individual has the same position in this year by pure chance [a,b,c]

Position in year 2008, given that residual expenses are positive [d]	Same position in year 2009	Same position in year 2010	Same position in year 2011
Top 1%	73.3	35.8	23.7
Top 5%	14.7	8.9	6.0
Top 10%	8.0	5.5	3.9
Top 20%	4.3	3.2	2.7
Top 50%	2.3	2.0	1.8
Bottom 50%	2.0	1.8	1.6

Footnotes Table 6.4:

a. Individual's position in a year was based on the distribution of residual expenses, given that residual expenses were positive. Residual expenses were based on the Dutch RE-model of 2013.

b. The statistics were based on insured who were enrolled in each year over the time period 2008 to 2011. Individuals who were not enrolled in each of the four years were excluded from this analysis due to missing information for one or more years; e.g. deceased individuals or new borns.

c. The statistics in this table were calculated as follows: the probability that an individual occurs in each specific position was divided by the probability that an individual can be in this position by pure chance. An example of interpreting the statistics in this table is: among those individuals in the top 1% in year 2008, the probability of staying in the top 1% in year 2009 is 73.3 times higher than can be expected by pure chance. In 2011, this probability is still 23.7 times higher than can be expected, indicating that there is some persistence in residual expenses. In other words, for individuals who are in the top 1% of positive residual expenses in a year it is likely that they stay in the top 1% in a following year.

d. Statistics with in the first column the position in year 2009 and 2010 are not presented here because these tables provide a similar pattern in persistence in residual expenses. Statistics of individuals who did not stay in the same position but for whom the position increase or decrease in a following year are also not presented here. These statistics show that the position of individuals may change over time; however, the probability that individuals with positive residual expenses have positive residual expenses in a following year is still larger than can be expected by pure chance. A table with all statistics is presented in Appendix 6.1.

is on the one hand some regression towards the mean (Beck & Zweifel, 1998; Beck et al., 2010; Welch, 1985), but on the other hand some persistence in residual expenses, implying that there are groups for whom the Dutch RE-model of 2013 under-compensates insurers over several years.

§ 6.3.2 Cost patterns of individuals with persistent under-compensations

Given the information in Table 6.4, many group definitions for those individuals with persistent under-compensations are possible. For a detailed analysis of cost patterns and risk characteristics of individuals who exhibit persistent under-compensations we focus on two groups, which are "individuals with positive residual expenses in each of three consecutive years" (*definition type 1*) and "individuals in the top 50% of those individuals with positive residual expenses in each of three consecutive years" (*definition type 2*). Since this choice is somewhat arbitrary, we define several alternative groups, which are "individuals with positive residual expenses in at least one year over three consecutive years" (*definition type 3*), "individuals with positive residual expenses in two consecutive years over a three year time period " (*definition type 4*), and "individuals in the top 1%, 5% or 10% of those individuals

with positive residual expenses in each of three consecutive years" (*definitions type 5, 6, and 7, respectively*). Table 6.5 presents the sizes of these groups in the population in 2011 and average residual expenses for these groups in the years 2008 to 2011.

Table 6.5 shows that 3.59% of the population in 2011 is under-compensated in each of the three previous years (i.e. definition type 1), which is significantly larger than the percentage of individuals that can be expected in this group by pure chance (~1.2%)[11]. Average residual expenses for this group in the years 2008 to 2010 are € ~3,200 to € ~3,700, but average residual expenses substantially reduce to € 943 in year 2011. Reasons for this significant reduction may be (a combination of) the effect of MHC-groups in the Dutch RE-model of 2013 that is estimated on the data from 2011, a reduction in observed expenses over time, and the occurrence of health episodes with a time duration of three years that were diagnosed in 2008. The MHC-groups adjust for high observed expenses in the three previous years, which for the RE-model that is estimated on data from 2011 exactly matches the time period that is used to identify the persistently under-compensated groups. Consequently,

Table 6.5: Pattern in average residual expenses over a four-year time period for individuals with persistent under-compensations under the Dutch RE-model of 2013

Group definitions [a]	**Percentage of total population in year 2011**	**Average residual expenses, in €'s** [b]			
		Year 2008	**Year 2009**	**Year 2010**	**Year 2011**
Type 1: Positive residual expenses in three consecutive years	3.59	3,254	3,427	3,701	943
Type 2: Top 50% of positive residual expenses in each of three consecutive years	0.74	9,342	9,130	9,188	1,389
Type 3: Positive residual expenses in at least one year over a three years	40.88	928	794	915	352
Type 4: Positive residual expenses in two consecutive years over three consecutive years	11.99	1,751	3,101	1,731	615
Type 5: Top 1% of positive residual expenses in each of three consecutive years	0.01	126,189	125,528	107,481	94,593
Type 6: Top 5% of positive residual expenses in each of three consecutive years	0.03	55,032	52,565	46,390	25,273
Type 7: Top 10% of positive residual expenses in each of three consecutive years	0.07	35,104	32,898	29,682	10,494

Footnotes Table 6.5:

a. The groups of individuals who are persistently under-compensated were identified by using administrative data from the years 2008 to 2010. Data from year 2011 was used here to examine whether the pattern in expenses that was found over three prior years is consistent with the one in this year (i.e. 2011 is the estimation-year for further analysis).

b. Residual expenses were calculated as: observed expenses minus predicted expenses of the Dutch RE-model of 2013. Expenses were annualized and weighted by the enrollment period in the respective year.

[11] Calculated as: 0.219*0.232*0.235 = ~1.2%. In the years 2008 to 2010, 21.9%, 23.2%, and 23.5% of the individuals had positive residual expenses, respectively.

residual expenses for those individuals are substantially reduced in 2011, since individuals with high observed expenses are likely to have high residual expenses (Table 6.3)[12,13].

Identifying the group of individuals in the top 50% of positive residual expenses in each of three consecutive years (i.e. definition type 2) result into selecting 0.74% of the population in 2011, which is also significantly larger than can be expected by pure chance (~0.15%)[14]. For this group, the same pattern in expenses can be observed as for the group that is identified by using definition type 1: average residual expenses are € ~9,000 in each of the years 2008 to 2010, but substantially reduce to € 1,389 in year 2011.

Alternative group definitions select different groups in the population: compared to definition types 1 and 2, types 3 and 4 select larger groups with lower average residual expenses in a year and types 5, 6, and 7 select smaller groups with higher average residual expenses in a year. Also for these alternative groups, average residual expenses in 2011 are (substantially) lower than in each of the three previous years.

§ 6.3.3 Risk characteristics of individuals with persistent under-compensations

Classification in a PCG, DCG, and/or MHC-group as indicators for individuals' health status

For the same groups as identified in the previous step, we analyze some risk characteristics of these individuals. Table 6.6 is a contingency table with those who were classified in a PCG, DCG, and/or MHC-group and those who were classified in none of these risk adjusters (i.e. the columns) to those who were persistently under-compensated and those who were not, according to definition types 1 and 2 (i.e. the rows). This table provides insight into how many individuals were classified in each group and how large average observed expenses and residual expenses are for each of these groups. A detailed analysis of the prevalence of PCGs, multiple PCGs, DCGs, MHC-groups, and a combination of these risk adjusters can be found in Appendix 6.2.

A first interesting result in Table 6.6 is that individuals with persistent under-compensations differ markedly from the total population in 2011. Among these individuals there is a higher prevalence of classification in a PCG, DCG, and/or MHC-group, compared to the

[12] When MHC-groups were *not* included in the RE-model, the reduction in average residual expenses from 2010 to 2011 is smaller than when MHC-groups were included in the RE-model.

[13] In an additional analysis, we examined the mortality rate among the individuals in the persistently under-compensated group in 2011. Mortality was approximated by selecting individuals with an age of 65 or older and being partly enrolled in 2011. This analysis showed that the mortality rate is not substantially higher among the individuals in the persistently under-compensated group than in the population.

[14] Calculated as: 0.109*0.116*0.118 = ~0.15%. In 2008 to 2010, 10.9%, 11.6%, and 11.8% of the individuals were in the top 50% of positive residual expenses in each year, respectively.

population: ~45%[15] of the individuals according to definition type 1 and ~85%[16] according to definition type 2 versus ~22% for the population. Further, individuals with persistent under-compensations *and* a PCG, DCG, and/or MHC-group have higher average observed expenses than the total group of individuals with a PCG, DCG, and/or MHC-group in the

Table 6.6: Risk characteristics of the persistently under-compensated groups on the total administrative dataset from 2011 (N = ~16.7 million)

			A PCG, DCG, and/or MHC-group [c]		Total
			Yes	No	
Type 1: Positive residual expenses in each of three consecutive years [a]	**Yes**	Percentage of the population	1.65%	1.94%	3.59%
		Average observed expenses [d], in €'s	8,041	1,817	4,674
		Average residual expenses [e], in €'s	881	996	943
	No	Percentage of the population	20.29%	76.12%	96.41%
		Average observed expenses [d], in €'s	4,337	968	1,677
		Average residual expenses [e] , in €'s	20	-50	-35
Type 2: Top 50% of positive residual expenses in each of three consecutive years [b]	**Yes**	Percentage of the population	0.63%	0.11%	0.74%
		Average observed expenses [d], in €'s	12,522	2,502	11,309
		Average residual expenses [e], in €'s	1,338	1,684	1,389
	No	Percentage of the population	21.30%	77.95%	99.25%
		Average observed expenses [d], in €'s	4,380	987	1,715
		Average residual expenses [e], in €'s	47	-26	-10
Total		Percentage of the population	21.94%	78.06%	100%
		Average observed expenses [d], in €'s	4,615	989	1,785
		Average residual expenses [e], in €'s	84	-24	0

Footnotes Table 6.6:

a. An individual was classified in this group if he/she had positive residual expenses (= observed expenses minus RE-predicted expenses) in each year over the time period 2008 to 2010 under the Dutch RE-model of 2013. All remaining individuals in the population in 2011 – i.e. those who did not have positive residual expenses in three consecutive prior years plus those for whom prior years' information was lacking (e.g. new borns) – were assigned to the complementary group.
b. An individual was classified in this group if he/she was in the top 50% of residual expenses, given that residual expenses were positive, in each year over the time period 2008 to 2010 under the Dutch RE-model of 2013. All remaining individuals in the population in year 2011 – i.e. those who were not in the 50 % of positive residual expenses in a year over three consecutive years plus those for whom prior years' information was lacking (e.g. new borns) – were assigned to the complementary group.
c. An individual was classified in the group 'yes' if he/she were classified in a PCG, DCG, and/or a MHC-group in 2011 and in the complementary group 'no' if he/she were not classified in a PCG, *and* a DCG, *and* a MHC-group. PCG: Pharmacy-based Cost Groups, individuals with multiple PCGs were counted only once in this calculation; DCG: Diagnostic Cost Groups; MHC-group: Multiple-year High Cost Groups.
d. Total observed expenses were the sum of expenses on services included in the Dutch benefit package in 2011.
e. Residual expenses were calculated as: observed expenses minus predicted expenses of the Dutch RE-model of 2013. Expenses were annualized and weighted by the enrollment period.

[15] Percentage calculated as: (1.65/3.59) * 100% = ~45%.

[16] Percentage calculated as: (0.63/0.74) * 100% = ~85%.

population: € 8,041 according to definition type 1 and € 12,522 according to definition type 2, versus € 4,615 for the group in the population, indicating that individuals with persistent under-compensations are the relatively high-risk individuals within the group of individuals with a PCG, DCG, and/or MHC-group in the population. Further detailed analysis of the prevalence of a PCG, DCG, and MHC-group indicates that a relatively large proportion of the individuals with persistent under-compensations are classified to a PCG *and* a DCG *and* a MHC-group (Appendix 6.2).

A second interesting result in Table 6.6 is that there are also individuals with persistent under-compensations who were *not* classified in a PCG, DCG, and MHC-group. For these individuals, average residual expenses are higher than for those who were classified in a PCG, DCG, and/or MHC-group, while observed expenses are lower. Apparently, these individuals are not identified with the (morbidity-based) risk adjusters in the Dutch RE-model of 2013. As a result, these individuals are classified in the reference groups of the PCGs, DCGs, and MHC-groups, resulting into a (substantially) lower risk-adjusted payment than individuals who are classified in a PCG, DCG, and MHC-group. Consequently, this may lead to under-compensations when these individuals are the relatively high-cost individuals within the reference group.

A detailed analysis of the risk characteristics of alternative group definitions shows that the risk characteristics of the persistently under-compensated group somewhat change when other definitions are used (Appendix 6.2). However, for all alternative group definitions the percentage of individuals classified in a PCG, DCG, and/or MHC-group, and a combination of these risk adjusters is higher than those in population in 2011, indicating that individuals with persistent under-compensations are relatively unhealthy.

Self-reported long-term diseases

In addition to the previous analysis, we examine the risk characteristics of the persistently under-compensated groups in the survey sample. Table 6.7 is a contingency table of those who reported a long-term disease and those who did not reported a disease (i.e. the columns) to those individuals who were persistently under-compensated and those who were not, according to definition types 1 and 2 (i.e. the rows).

Table 6.7 shows that among the individuals with persistent under-compensations there is a higher prevalence of long-term diseases, compared to the population: ~60%[17] according to definition type 1 and ~75%[18] according to definition type 2, versus ~32% in the population. These groups have average observed expenses of € 7,926 and € 17,043 according to definition types 1 and 2, respectively, which is far above average observed expenses of € 3,480 in

[17] Percentage calculated as: (2.28/3.82) * 100% = ~60%.

[18] Percentage calculated as: (0.62/0.83) * 100% = ~75%.

Table 6.7: Risk characteristics of the persistently under-compensated group on the survey sample from year 2010 (N = 16,141)

			A self-reported long-term disease [c]		Total
			Yes	No	
Type 1: Positive residual expenses in each of three consecutive years [a]	**Yes**	Percentage of the population	2.28%	1.54%	3.82%
		Average observed expenses [d], in €'s	7,926	2,018	5,541
		Average residual expenses Model 0 [e,f], in €'s	2,321***	244	1,483***
	No	Percentage of the population	29.25%	66.93%	96.18%
		Average observed expenses [d], in €'s	3,133	952	1,616
		Average residual expenses Model 0 [e,f], in €'s	182**	-164***	-59*
Type 2: Top 50% of positive residual expenses in each of three consecutive years [b]	**Yes**	Percentage of the population	0.62%	0.21%	0.83%
		Average observed expenses [d], in €'s	17,043	3,816	13,746
		Average residual expenses Model 0 [e,f], in €'s	3,453**	-1,511	2,216*
	No	Percentage of the population	30.90%	68.26%	99.17%
		Average observed expenses [d], in €'s	3,206	968	1,665
		Average residual expenses Model 0 [e,f], in €'s	274***	-151***	-19
Total		Percentage of the population	31.53%	68.47%	100%
		Average observed expenses [d], in €'s	3,480	976	1,766
		Average residual expenses Model 0 [e,f], in €'s	337***	-155***	0

Footnotes Table 6.7:

a. An individual was classified in this group if he/she had positive residual expenses (= observed expenses minus RE-predicted expenses) in each year over the time period 2008 to 2010 under the Dutch RE-model of 2013. All remaining individuals in the population in year 2011 – i.e. those who did not have positive residual expenses in three consecutive prior years plus those for whom prior years' information was lacking (e.g. new borns) – were assigned to the complementary group.

b. An individual was classified in this group if he/she was in the top 50% of residual expenses, given that residual expenses are positive, in each year over the time period 2008 to 2010 under the Dutch RE-model of 2013. All remaining individuals in the population in year 2011 – i.e. those who were not in the 50 % of positive residual expenses in a year over three consecutive years plus those for whom prior years' information was lacking (e.g. new borns) – were assigned to the complementary group.

c. An individual was classified in the group 'yes' if he/she had a answered at least one of the following questions with a 'yes': 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 regularly periods of dizziness 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? If all these questions were answered with a 'no', an individual was classified in the complementary group. Individuals with a missing value for one of the aforementioned questions were assumed to have not the self-reported disease that was asked for.

d. Total observed expenses were the sum of expenses on services included in the Dutch benefit package in 2011.

e. Residual expenses were calculated as: observed expenses minus predicted expenses of the Dutch RE-model of 2013. Expenses were annualized and weighted by the enrollment period. Residual expenses were calibrated: individuals' residual expenses per RE-model were raised by a factor equaling average RE-predicted expenses in the survey sample divided by average observed expenses in the survey sample. After this calibration, average residual expenses per RE-model were zero in the survey sample, just as is the case in the population.

f. Since we used a sample of the total population, we examined the statistical significance of our results: ***: statistically significantly different from zero with a P-value ≤ 0.01, **: statistically significantly different from zero with a P-value ≤ 0.05, *: statistically significantly different from zero with a P-value ≤ 0.10, based on an one-sample two-sided T-test.

the population with a long-term disease. Further, these groups have high average residual expenses of € 2,321 and € 3,453 according to definition types 1 and 2, respectively (both statistically significantly different from zero at 1%). These statistics indicate that individuals with persistent under-compensations are the relatively high risks among those with a long-term disease. An additional analysis shows that more than half of all individuals with persistent under-compensations have multiple long-term diseases. Table 6.7, however, also shows that some individuals with persistent under-compensations reported no long-term disease. For these groups, average residual expenses are substantially lower than those with persistent under-compensations *and* a long-term disease (not statistically significantly different from zero at 10%).

The above statistics show that individuals with persistent under-compensations under the Dutch RE-model of 2013 are relatively unhealthy and often have multiple long-term diseases. Based on these results, we can address two potential reasons why the Dutch RE-model of 2013 does not predict expenses adequately for these individuals: (i), the morbidity-based risk adjusters are heterogeneous with respect to expected expenses, whereby the individuals with persistent under-compensations are the relatively high-cost individuals within a risk class; and (ii), some individuals are *not* classified in a PCG, DCG, and/or MHC-group. A reason for this may be restrictions on classification in these morbidity-based risk adjusters; e.g. an individual should have used more than 180 Described Daily Dosages of specific drugs in order to be classified in a PCG.

§ 6.3.4 Predictive power of a risk adjuster or interaction terms for the persistently under-compensated group

Model's predictive performance for the full sample

The information from the previous sections can be used to define an explicit risk adjuster or interaction terms for the persistently under-compensated group in the population in 2011, aiming to examine to what extent the predictive performance of the Dutch RE-model of 2013 can be improved. As shown in Table 6.8, extending Model 0 (i.e. the Dutch RE-model of 2013) with a risk adjuster for the persistently under-compensated group increases model's predictive performance. Model 0 has a R^2 of 24.22%, a CPM of 24.88% and a MAPE of € 1,570. According to definition type 1 for the persistently under-compensated group, the R^2 increases by 0.10 percentage points, the CPM increases by 0.47 percentage points, and the MAPE reduces by € 9. Using definition type 2 instead of type 1 to define a risk adjuster leads to a smaller improvement of model's predictive performance: plus 0.05 percentage points in R^2, plus 0.08 percentage points in CPM, and minus € 1 in MAPE, compared to Model 0. The coefficient of the risk adjuster according to definition types 1 and 2 is € 1,041 and € 1,594, respectively, implying that an insurer would receive a substantially higher payment for individuals with persistent under-compensations compared to those who have not.

Table 6.8: Predictive performance of the estimated RE-models on the full validation sample (*N* = ~8.4 million) [a]

	Adj.-R^2, in % [b]	CPM, in % [c]	MAPE, in €'s [d]
Model 0 (Dutch RE-model 2013)	24.22	24.88	1,570
Model 1 (Model 0 + a dummy variable for the persistently under-compensated group according to *definition type 1*) [e]	24.32	25.35	1,561
Model 2 (Model 0 + a dummy variable for the persistently under-compensated group according to *definition type 2*) [f]	24.27	24.96	1,569
Model 3 (Model 0 + interaction terms for the persistently under-compensated group according to *definition type 1*) [e, g]	25.73	25.54	1,557
Model 4 (Model 0 + interaction terms for the persistently under-compensated group according to *definition type 2*) [f, g]	25.86	25.14	1,565

Footnotes Table 6.8:

a. All RE-models were estimated on estimation-sample of the administrative dataset and expenses were predicted on the validation-sample. All RE-models were evaluated on the validation-sample.
b. Adj.-R^2 = adjusted-R-squared. The adj.-R^2 was 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.
c. CPM = Cumming's Prediction Measure. The CPM was calculated as one minus the ratio of the MAPE to the mean absolute difference between observed expenses and average observed expenses.
d. MAPE = Mean Absolute Prediction Error. The MAPE was calculated as the absolute difference between predicted expenses and observed expenses. The MAPE was rounded to the nearest €.
e. Group definition type 1: an individual was classified in the persistently under-compensated group if he/she had positive residual expenses in each of the three previous years under the Dutch RE-model of 2013. All remaining individuals in the population were classified in the complementary group.
f. Group definition type 2: an individual was classified in the persistently under-compensated group if he/she was in the top 50% of residual expenses, given that residual expenses were positive, in each of the three previous years under the Dutch RE-model of 2013. All remaining individuals in the population were classified in the complementary group.
g. Interaction terms meant inclusion of an interaction term between each risk adjuster in the Dutch RE-model of 2013 and a dummy variable for the persistently under-compensated group that was identified.

Furthermore, using interaction terms lead to a larger improvement of model's predictive performance than using one extra risk adjuster for the persistently under-compensated group. For Model 3 compared to Model 0, the R^2 increases by 1.51 percentage points, the CPM increases by 0.66 percentage points, and the MAPE reduces by € 13. For Model 4 compared to Model 0, the R^2 increases by 1.64 percentage points, the CPM increases by 0.26 percentage points, and the MAPE reduces by € 5.

The above statistics show that the increase in percentage points in R^2 is smaller than the increase in percentage points in CPM when a risk adjuster is included, while the opposite holds when interaction terms are included. This finding indicates that interaction terms are better able to predict expenses for some individuals in the right-tail of the cost distribution than an extra risk adjuster, because the R^2 weighs large residual expenses more heavily than small residual expenses, while the CPM weighs them equally.

Sensitivity analysis

To indicate to what extent our conclusions about model's predictive performance change when an alternative group definition is used to define an extra risk adjuster, we estimated

five additional RE-models according to the aforementioned alternative group definitions (i.e. definition types 3 to 7).

Estimation of the alternative RE-models shows that the R^2 ranges from 24.38% to 26.09%, the CPM from 24.91% to 26.12%, and the MAPE from € 1,545 to € 1,570 (see Appendix 6.3 for detailed statistics per RE-model). Of these alternative RE-models, the model with a risk adjuster based on the group of individuals with positive residual expenses in at least one year over three consecutive prior years (i.e. definition type 3) has the highest predictive performance in terms of CPM and MAPE for the full sample (not in terms of R^2). A reason for this may be that the extra risk adjuster in this RE-model may specifically identify individuals with a health episode in 2010 (and not in 2009 or 2008), which are likely to have high expenses in the next year, in this case the estimation-year. If two consecutive years are incorporated for defining the persistently under-compensated group (i.e. definition type 4), the predictive power of the risk adjuster reduces compared to using definition type 3; however, model's predictive performance is still somewhat higher than inclusion of a risk adjuster according to definition types 1 or 2 that are based on three consecutive previous years. A reason for this may be that fluctuations due to short-term health episodes or accidents at the end of the calendar year still can play a role, because two instead of three consecutive years are used. Using the top 1% of positive residual expenses in each year over a three-year time period results into the largest R^2 of all RE-models (not the largest CPM), indicating that this model specifically predicts expenses in the right-tail of the cost distribution. To conclude, this sensitivity analysis shows that using an alternative definition for the persistently under-compensated group changes the extent to which Model 0 can be improved in terms of R^2, CPM and MAPE on the full sample; however, all group definitions show that there is a potential for improving the predictive performance of the Dutch RE-model of 2013.

Model's predictive performance for selective groups

Out of all 46 pre-defined selective groups, there are 16 groups for which average residual expenses of one or more estimated RE-models are statistically significantly different from zero (Table 6.9). For the other 30 groups, average residual expenses of all RE-models are *not* statistically significantly different from zero and therefore, are not presented here (see Appendix 6.4). Note that positive average residual expenses for a group imply negative residual expenses for the complementary group and vice versa. Furthermore, individuals can be classified in multiple groups. Table 6.9 presents average residual expenses per group for Model 0. For Models 1 to 4, the differences between average residual expenses for these models and average residual expenses for Model 0 are presented in order to indicate the extent to which inclusion of a risk adjuster or interaction terms improve the prediction of expenses for each group. The statistical significance of average residual expenses for a group (and *not* the difference in average residual expenses) is indicated by asterisks.

Table 6.9: Predictive performance of the estimated RE-models for selective groups with average residual expenses that are statistically significantly different from zero, on the survey sample from year 2010 (N = 16,141) [a, b]

Groups	Size, in %	Average observed expenses, in €'s	Average residual expenses in year 2011, in €'s				
			Model 0 [d]	(Model 1 – 4 minus Model 0) [c]			
				Model 1 [e]	Model 2 [f]	Model 3 [g]	Model 4 [h]
General health status (all respondents)							
General health status is poor	19.0	4,279	375**	-23**	-7**	-5**	10**
At least one long-term disease	31.6	3,480	337**	-24**	-43**	-57**	-43**
Functional disabilities (age ≥ 12 years)							
OECD limitations in hearing	2.9	5,133	1,208*	31*	15*	14*	-4*
OECD limitations in seeing	6.1	3,883	728*	-1*	-3*	36*	28*
Scores on SF-12 (age ≥ 12 years)							
The lowest score on physical health scales	9.5	5,707	845*	-6*	5*	41*	47*
A low score on physical health scales	19.0	4,476	672**	-7**	5**	8**	25**
Presence of disease or disorder (age ≥ 12 years)							
Serious bowel disorders, longer than 3 months	3.4	4,412	820*	-7*	7*	-24*	-14*
Serious / persistent back problems or pain	10.6	3,488	363*	-5	-5	-22	-21
Co-morbidity (age ≥ 12 years)							
Three or more self-reported diseases or (chronic) disorder	17.5	4,212	337	-11	4*	14*	32*
Health care utilization (all respondents)							
Contact general practitioner in the past year	72.0	2,068	81	-7	-1	-2	3*
Contact medical specialist in the past year	37.9	3,114	327**	-29**	-7**	-18**	3**
Hospitalization in the past year	6.6	5,773	591*	-31*	-25*	45*	51*
Contact physiotherapist in the past year	21.8	2,934	324**	-31**	-10**	-27**	-11**
Prescribed drugs use in the past 14 days	35.7	3,133	188**	-21*	-4*	-24*	-9*
Use of durable medical equipment	7.2	5,094	621	12*	27*	-10	19*
Contact with home nurse (care) in the past year	1.4	9,336	1,402*	15*	34*	-2*	2*

Footnotes Table 6.9:

a. Residual expenses were calculated as: observed expenses minus predicted expenses of the Dutch RE-model of 2013. Expenses were annualized and weighted by the enrollment period. Residual expenses were calibrated: individuals' residual expenses per RE-model were raised by a factor equaling average RE-predicted expenses in the survey sample divided by average observed expenses in the survey sample. After this calibration, average residual expenses per RE-model were zero in the survey sample, just as is the case in the population. For Model 0, average residual expenses per group are presented. For Models 1 to 4, the difference in average residual expenses per group between these models and Model 0 are presented.

b. **: Statistically significantly different from zero with a P-value ≤ 0.01; *: Statistically significantly different from zero with a P-value ≤ 0.05, based on an one-sample two-sided T-test. The statistical significance of Model 1 to 4 refers to the statistical significance of average residual expenses and *not* to the difference in average residual expenses compared to Model 0.

c. The statistics indicate the difference in average residual expenses for Model 1 to 4, respectively to Model 0. A negative value imply that average residual expenses of this model for this group are reduced and a positive value imply that average residual expenses are increased, compared to Model 0.

d. Model 0: the Dutch RE-model of 2013.

e. Model 1: Model 0 + a dummy variable for the persistently under-compensated group according to definition type 1 (i.e. positive residual expenses in each of three consecutive years).

f. Model 2: Model 0 + a dummy variable for the persistently under-compensated group according to definition type 2 (i.e. top 50% of positive residual expenses in each of three consecutive years).

g. Model 3: Model 0 + interaction terms for the persistently under-compensated group according to definition type 1 (i.e. positive residual expenses in each of three consecutive years).

h. Model 4: Model 0 + interaction terms for the persistently under-compensated group according to definition type 2 (i.e. top 50% of positive residual expenses in each of three consecutive years).

Table 6.9 shows that Model 0 under-predicts expenses for several selective groups; e.g. individuals with a poor general health status, or at least one long-term disease, or a low score on physical health scales. Models 1 to 4 reduce average residual expenses for several groups, compared to Model 0; e.g. individuals with at least one long-term disease, or those with serious back problems or back pain, or those who contacted a physiotherapist in the past year. However, average residual expenses for these groups are still statistically significantly different from zero, implying that these extended RE-models also do not predict expenses adequately for these groups. In addition, Models 1 to 4 increase average residual expenses for other groups, compared to Model 0; e.g. individuals with limitations in hearing for Models 1 to 4, individuals who use durable medical equipment for Models 1, 2, and 4, and individuals with three or more self-reported diseases for Models 2 to 4. These results show that extending Model 0 with a risk adjuster or interaction terms for the persistently under-compensated group may improve the prediction of expenses for some selective groups, but may deteriorate the prediction of expenses for others. Further, these extended RE-models still result into under-compensations for several selective groups that are statistically significantly different from zero.

§ 6.4 CONCLUSIONS AND DISCUSSION

This study explores whether there is a group that is *persistently* under-compensated under a morbidity-based RE-model, which is the Dutch RE-model of 2013. A rich cross-sectional time series dataset covering almost the entire Dutch population (N = ~16 million) for the time period 2008 to 2011 is used combined with survey data from 2010 (N = 16,141).

A first conclusion of this study is that there is some persistence in residual expenses under the Dutch RE-model of 2013, implying that there are groups in the population who exhibit persistent under-compensations over a period of three years. The probability that individuals stay in the top percentiles of residual expenses declines over years, but this probability is still much larger than can be expected by pure chance. Further, the probability that individuals continue to have (high) positive residual expenses in a following year, whereby this probability is adjusted for the probability that individuals can be in this position by pure chance, increases when an individual is in a higher residual percentile in a prior year; e.g. the top 1% of positive residual expenses versus the top 50%. The risk characteristics of individuals who exhibit persistent under-compensations differ markedly from those of the population. Based on our analysis, these individuals have high above-average observed expenses and are relatively unhealthy compared to the population. Further, most of them have multiple long-term diseases. Based on these findings, we can address two potential reasons why the Dutch RE-model of 2013 persistently under-compensates insurers for a sizeable group in the population, which are: (i), the morbidity-based risk adjusters in this

RE-model are heterogeneous with respect to expected expenses, whereby the individuals with persistent under-compensations are the relatively high-cost individuals within a risk class; (ii), some individuals with persistent under-compensations are *not* classified in any of the morbidity-based risk adjusters, which are in this case the pharmacy-based cost groups (PCG), diagnostic cost group (DCG), and/or multiple-year high costs groups (MHC-groups).

A second conclusion of this study is that extending the Dutch RE-model of 2013 with an explicit risk adjuster or interaction terms defining the persistently under-compensated group increases model's predictive performance for the full sample and for some selective groups of interest. However, this study also shows that inclusion of a risk adjuster or interaction terms may deteriorate the prediction of expenses on other groups. Consequently, inclusion of a risk adjuster or interaction terms mitigates financial incentives for risk selection for some groups but does not eliminate them.

In this study, two definitions of the persistently under-compensated group are examined in detail, namely: "individuals with positive residual expenses in each of three consecutive years" and "individuals in the top 50% of positive residual expenses in each of three consecutive years". A sensitivity analysis shows that using an alternative group definition changes our conclusions about the size and the risk characteristics of the persistently under-compensated group. Identifying individuals in the top 1%, 5%, or 10% of positive residual expenses in each of three consecutive years results into selecting a smaller group in the population with higher average residual expenses, compared to the two aforementioned definitions that were examined in detail. In addition, there is a higher prevalence of classification in a PCG, DCG, and/or MHC-groups among these individuals. Identifying individuals who have positive residual expenses in at least two out of the three preceding years results into selecting a larger group in the population with smaller average residual expenses and a lower prevalence of classification in a PCG, DCG, and/or MHC-group, compared to the two aforementioned definitions that were examined in detail. Furthermore, an alternative group definition changes the extent to which model's predictive performance can be improved when a risk adjuster defining the persistently under-compensated group is included in the Dutch RE-model of 2013. Based on our analysis, model's predictive performance can be improved by 0.05 to 1.87 percentages points in the R-squared (R^2), 0.03 to 1.14 percentage points in the Cumming's Prediction Measure (CPM), and € 1 to € 25 in the Mean Absolute Prediction Error (MAPE), depending on the exact group definition that is used for defining the risk adjuster.

Since our analysis has exclusively explored persistent under-compensations under the Dutch RE-model of 2013, we cannot conclude whether there are individuals with persistent under-compensations under other morbidity-based RE-models. It is expected that for an RE-model with a lower predictive performance than the Dutch RE-model of 2013 the persistently under-compensated group is larger than the group that is identified in this

study. Moreover, the inclusion of a risk adjuster or interaction terms defining this group would probably lead to a larger improvement in the model's predictive performance. In principle, the method that is applied in this study to explore whether there are individuals with persistent under-compensations is generally applicable to any RE-model.

§ 6.4.1 Health-policy implications

Individuals with persistent under-compensations are vulnerable to risk selection. Given our empirical analysis, policymakers can mitigate financial incentives for risk selection for some selective groups of interest by extending the RE-model with a risk adjuster or interaction terms defining the persistently under-compensated group under this RE-model. However, our analysis shows that it takes a relatively large amount of effort in terms of applying advanced methods on large datasets in order to improve the prediction of expenses for a relatively small group of individuals who exhibit persistent under-compensations. Although these individuals can be identified and we know their risk characteristics, it is difficult to compensate insurers adequately for predictable variation in individuals' healthcare expenses. Since individuals with persistent under-compensations have far above-average expenses, it is needed to redistribute a large amount of money from other individuals in the population to individuals with persistent under-compensations in order to improve the prediction of expenses for them. Hence, to accomplish a better prediction for a small group in the population, the prediction of expenses for other groups may deteriorate. This concept can be observed by the relatively large increase in the R^2 when a risk adjuster or interaction terms defining the persistently under-compensated group is included in the RE-model, while there is only a slightly increase in the CPM. The R^2 weighs large residual expenses more heavily than small residual expenses and so, this measure is more sensitive to outlier observations compared to the CPM, which weighs residual expenses equally. Further, this concept is also reflected in the results of the group-level analysis. Average residual expenses for selective groups of interest do not statistically significantly reduce as a result of including a risk adjuster or interaction terms defining the persistently under-compensated group, because a better prediction of expenses for some individuals may be cancelled out by a slightly worse prediction for other individuals within the same group.

This study has shown that extending the RE-model with a risk adjuster or interaction terms defining the persistently under-compensated group may improve model's predictive performance. It should be noted, however, that the predictive performance of an RE-model is not the only relevant evaluation criterion for RE-models. To policymakers more criteria may play a role in addition to model's predictive performance when deciding on the design of the RE-model, such as the vulnerability to manipulation and appropriateness of incentives for efficiency (van de Ven & Ellis, 2000). When a risk adjuster or interaction terms defining the persistently under-compensated group are included in the RE-model, an insurer may receive higher risk-adjusted payments for individuals with persistent under-

compensations compared to those who have not. Consequently, it should not be easy to manipulate whether an individual is classified in the persistently under-compensated group. It is likely that using a definition like "individuals being in the top 50% of those individuals with positive residual expenses in each of three consecutive years" may be difficult to manipulate, because an individual should have high residual expenses in each year over a three-year time period. Such a definition is similar to the MHC-groups that are used in the Netherlands. Further, appropriateness of incentives for efficiency may play a role to policymakers. In this study, the persistently under-compensated group is based on prior years' residual expenses, which is the difference between observed expenses and RE-predicted expenses in a year, whereby we assume that residual expenses are completely determined by risk factors for which the regulator desires compensation, such as individuals' health status (i.e. the S-type cost variation). We did not incorporate that observed expenses also include random variation and variation for which the regulator does *not* desire compensation, such as practice variation or inefficiencies (i.e. the N-type cost variation). It is likely that residual expenses are determined by a combination of S-type and N-type risk factors. In practice, however, it is difficult to disentangle S-type and N-type cost variation and so, to define a risk adjuster or interactions terms based on prior years' residual expenses that completely remove financial incentives for risk selection, while respecting at the same time appropriateness of incentives for efficiency (Schokkaert & van de Voorde, 2004, 2006, 2009). Consequently, inclusion of a risk adjuster or interaction terms based on prior years' residual expenses may involve making a trade-off between mitigating financial incentives for risk selection and providing appropriate incentives for efficiency. Policymakers should decide how to weigh these two criteria. As long as there are no better alternatives to identify the persistently under-compensated group than using prior years' residual expenses (e.g. via risk adjusters that identify selective patient groups, such as individuals with specific types of cancer or rare diseases, which are expected to be persistently under-compensated because of their high observed expenses each year) and policymakers consider the financial incentives for risk selection for individuals with persistent under-compensations too large, they can include a risk adjuster or interaction terms based on prior years' residual expenses.

This study shows that financial incentives for risk selection for some selective groups can be mitigated but they cannot be eliminated. There may still be (significant) under- and over-compensations for some selective groups, which emphasizes the importance of further research on how to improve the predictive performance for sophisticated morbidity-based RE-models. For doing so, it may be relevant to start with exploring the risk characteristics of individuals with expenses in the (far) right-tail of the residual distribution, in a much more detail as done in this study. Such research may provide valuable insight into how to improve the prediction of expenses for these individuals.

§ 6.4.2 Study limitations

Our empirical findings are based on the Dutch RE-model of 2013, which includes a risk adjuster based on high observed healthcare expenses in each of the three previous years: the MHC-groups. The MHC-groups in the RE-model that is estimated on data from 2011 exactly match the time period that is used to identify individuals who exhibit persistent under-compensations in each of the three previous years. As a result of this, average residual expenses for the persistently under-compensated groups (substantially) reduce in this year because the MHC-groups to some extent improve the prediction for these groups. Exploring persistent under-compensations under an RE-model *without* MHC-groups may lead to the identification of a group with larger average residual expenses in each of the three previous years and the estimation-year than the groups that are identified in this study. Consequently, the extent to which an RE-model *without* MHC-groups can be improved by inclusion of a risk adjuster or interaction terms defining the persistently under-compensated group under this RE-model may be larger than found in this study.

In this study, the total population of insured is divided into the persistently under-compensated group and the complementary group. In practice, the population can also be subdivided in more than two distinctive groups; for example, a group with low under-compensations, a group with middle under-compensations, and a group with high under-compensations. More distinctive groups may improve model's predictive performance because more refined risk classes can be defined when this information is used to define a risk adjuster or interaction terms in the RE-model; however, it may also make model estimation more complex. Further research waits to investigate to what extent further refinement of defining risk groups in the population would improve model's predictive performance and whether this outweighs the increased complexity of the model. Such research should also examine the stability of the coefficients of the RE-model over time, because using more refined risk classes in the RE-model may increase the chance of random error.

Appendices
Chapter 6

APPENDIX 6.1: PERSISTENCE IN RESIDUAL EXPENSES OVER A FOUR-YEAR TIME PERIOD

Table A.6.1: Ratios of the actual probability that an individual is in a specific position in a following year to the expected probability that an individual is in this position in this year by pure chance [a,b,c]

	Position in year 2009						**Position in year 2010**						**Position in year 2011**					
Position in year 2008	*Top 1%*	*Top 5%*	*Top 10%*	*Top 20%*	*Top 50%*	*Bot. 50%*	*Top 1%*	*Top 5%*	*Top 10%*	*Top 20%*	*Top 50%*	*Bot. 50%*	*Top 1%*	*Top 5%*	*Top 10%*	*Top 20%*	*Top 50%*	*Bot. 50%*
Top 1%	**73.3**	26.1	15.8	9.2	4.3	0.5	**35.8**	13.9	8.9	5.6	3.0	0.6	**23.7**	10.4	6.9	4.5	2.6	0.7
Top 5%	26.5	**14.7**	10.3	6.7	3.4	0.5	13.6	**8.9**	6.7	4.5	2.6	0.6	9.0	**6.0**	4.7	3.4	2.2	0.7
Top 10%	15.4	10.1	**8.0**	5.6	3.1	0.6	8.6	6.6	**5.5**	4.0	2.4	0.7	5.9	4.6	**3.9**	3.0	2.1	0.8
Top 20%	8.7	6.6	5.6	**4.3**	2.8	0.8	5.3	4.5	4.0	**3.2**	2.2	0.8	3.9	3.4	3.1	**2.7**	2.0	0.9
Top 50%	4.1	3.5	3.3	2.9	**2.3**	1.2	2.8	2.6	2.5	2.3	**2.0**	1.1	2.3	2.2	2.1	2.0	**1.8**	1.1
Bottom 50%	0.7	0.9	1.0	1.2	1.5	**2.0**	0.8	0.9	1.0	1.1	1.3	**1.8**	0.8	0.9	1.0	1.1	1.3	**1.6**
Position in year 2009																		
Top 1%							**61.2**	24.7	15.4	9.3	4.5	0.4	**32.8**	14.9	9.9	6.3	3.4	0.6
Top 5%							22.2	**14.3**	10.5	6.9	3.6	0.5	11.6	**8.0**	6.1	4.4	2.6	0.7
Top 10%							13.3	9.9	**8.0**	5.7	3.2	0.6	7.4	5.7	**4.8**	3.7	2.4	0.8
Top 20%							7.8	6.4	5.5	**4.3**	2.8	0.8	4.6	3.9	3.5	**3.0**	2.2	0.9
Top 50%							3.8	3.4	3.2	2.8	**2.4**	1.2	2.5	2.4	2.3	2.2	**1.9**	1.2
Bottom 50%							0.7	0.8	0.9	1.1	1.4	**2.0**	0.8	0.8	0.9	1.0	1.3	**1.7**
Position in year 2010																		
Top 1%													**48.5**	21.6	14.0	8.6	4.3	0.5
Top 5%													18.1	**11.9**	8.9	6.1	3.4	0.6
Top 10%													11.3	8.4	**6.9**	5.1	3.1	0.7
Top 20%													6.8	5.6	4.9	**3.9**	2.7	0.9
Top 50%													3.5	3.1	3.0	2.7	**2.3**	1.2
Bottom 50%													0.6	0.7	0.9	1.0	1.3	**1.9**

Footnotes Table A.6.1:

a. The position per year was based on the distribution of residual expenses, given that residual expenses were positive.
b. The statistics were based on insured who were enrolled in each year over the time period 2008 to 2011. Individuals who were not enrolled in each of the four years were excluded from this analysis due to missing information for one or more years; e.g. deceased individuals or new borns.
c. The statistics in this table were calculated as follows: the probability that an individual occurs in each specific position was divided by the probability that an individual can be in this position by pure chance. An example of interpreting the statistics in this table is: among those individuals in the top 1% in 2008, the probability of being in the top 1% in 2009 is 73.3 times higher than the probability that individuals occur in this class by pure chance. In 2011, this probability is still 23.7 times higher than can be expected, indicating that there is some persistence in residual expenses. Thus, for individuals who are in the top 1% in a year it is likely that they stay in the top 1% in a following year.

APPENDIX 6.2: DETAILED RISK CHARACTERISTICS OF THE PERSISTENTLY UNDER-COMPENSATED GROUPS

Table A.6.2: Risk characteristics of individuals with persistent under-compensations under the Dutch RE-model of 2013, using the total administrative dataset in year 2011 (N = ~16.7 million) [a,b]

		Group size, in %	A PCG, in %	Multiple PCGs, in %	A DCG, in %	A MHC-group, in %	Co-morbidity, in % [d]	
							2	3
Total dataset	Records with prior's year information [c]	89.26	18.39	3.74	9.16	6.19	5.35	2.58
	Records with missing information from prior years (i.e. new borns)	10.74	8.28	1.20	4.40	2.33	2.18	0.80
Group definitions								
Type 1: Positive residual expenses in three consecutive years	Yes	3.59	25.97	5.51	21.50	30.02	13.25	9.17
	No	85.67	18.07	3.67	8.65	5.19	5.02	2.30
Type 2: Top 50% of positive residual expenses in each year over three consecutive years	Yes	0.74	42.92	13.31	42.34	78.35	27.60	25.40
	No	88.52	18.18	3.66	8.88	5.58	5.17	2.39
Alternative group definitions								
Type 3: Positive residual expenses in at least one year over a three-year time period	Yes	40.88	23.82	5.19	16.12	11.22	8.97	4.62
	No	48.38	13.79	2.52	3.28	1.93	2.30	0.86
Type 4: Positive residual expenses in two consecutive years over three years	Yes	11.99	26.95	5.99	20.06	22.84	12.03	8.13
	No	77.27	17.06	3.39	7.47	3.60	4.32	1.72
Type 5: Top 1% of positive residual expenses in each year over three consecutive years	Yes	0.01	40.75	17.04	70.87	98.90	49.16	30.68
	No	89.25	18.38	3.74	9.16	6.18	5.35	2.58
Type 6: Top 5% of positive residual expenses in each year over three consecutive years	Yes	0.03	56.14	25.68	68.52	90.39	35.57	41.42
	No	89.23	18.37	3.74	9.14	6.16	5.34	2.57
Type 7: Top 10% of positive residual expenses in each year over three consecutive years	Yes	0.07	59.47	25.24	67.75	85.36	32.24	42.47
	No	89.19	18.35	3.73	9.12	6.13	5.33	2.55

Footnotes Table A.6.2:

a. Statistics were weighted by the enrollment period in 2011.

b. For each group that was identified (i.e. groups in the first column) we calculated the percentage of individuals that were classified in a PCG, DCG, MHC-group, or a combination of these risk adjusters. Thus, the total group in the first column was used in the divisor for calculating the percentages.

c. The statistics of this group were used to compare the statistics of the group of individuals with persistent under-compensations (i.e. this group is the reference group).

d. Co-morbidity was defined as individuals that were classified in multiple morbidity-based risk adjusters, which were the PCGs, DCGs, and/or MHK-groups, whereby "PCG + DCG + MHK-group = 2" or "PCG + DCG + MHK-group = 3". Individuals with multiple PCGs were counted only once in determining co-morbidity. Note that individuals with *one* long-term disease can be classified to multiple morbidity-based risk adjusters.

APPENDIX 6.3: RESULTS OF THE SENSITIVITY ANALYSIS

Table A.6.3: Sensitivity analysis of the predictive performance of the Dutch RE-model of 2013 with an extra risk adjuster for the persistently under-compensated group according to alternative group definitions, by evaluating these RE-models on the full validation sample of the administrative dataset in year 2011 (*N* = ~8.4 million) [a]

RE-models with alternative group definitions	Adj. R^2, in % [b]	CPM, in % [c]	MAPE, in €'s [d]
Model 5 (Model 0 + a dummy variable for the persistently under-compensated group according to *definition type 3*: positive residual expenses in year 2008, or 2009, or 2010)	24.48	26.11	1,545
Model 6 (Model 0 + a dummy variable for the persistently under-compensated group according to *definition type 4*: positive residual expenses in two consecutive years over three years)	24.38	25.75	1,552
Model 7 (Model 0 + a dummy variable for the persistently under-compensated group according to *definition type 5*: top 1% of positive residual expenses in each of three consecutive years)	26.09	25.09	1,566
Model 8 (Model 0 + a dummy variable for the persistently under-compensated group according to *definition type 6*: top 5% of positive residual expenses in each of three consecutive years)	24.88	24.98	1,568
Model 9 (Model 0 + a dummy variable for the persistently under-compensated group according to *definition type 7*: top 10% of positive residual expenses in each of three consecutive years)	24.49	24.91	1,570

Footnotes Table A.6.3:

a. All RE-models were estimated on estimation-sample of the administrative dataset and expenses were predicted on the validation-sample. The statistics in this table were calculated on the validation-sample.

b. Adj.-R^2 = adjusted-R-squared. The adj.-R^2 was 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.

c. CPM = Cumming's Prediction Measure. The CPM was calculated as one minus the ratio of the MAPE to the mean absolute difference between observed expenses and average observed expenses.

d. MAPE = Mean Absolute Prediction Error. The MAPE was calculated as the absolute difference between predicted expenses and observed expenses. The MAPE was rounded to the nearest €.

APPENDIX 6.4: THE PREDICTIVE PERFORMANCE FOR SOME EVALUATION-GROUPS

Table A.6.4: Predictive performance of the estimated RE-models for selective groups with average residual expenses that are *NOT* statistically significantly different from zero, on the survey sample from year 2010 (N = 16,141) [a]

Groups	Size, in %	Average observed expenses, in €'s	Average residual expenses in year 2011, in €'s				
			Model 0 [b]	(Model 1-4 minus Model 0)			
				Model 1 [c]	Model 2 [d]	Model 3 [e]	Model 4 [f]
General health status (all respondents)							
Obesity	8.7	3,169	219	-10	-1	-33	-21
Functional disabilities (age ≥ 12 years)							
OECD limitations in talking	0.3	7,534	1,579	-5	86	341	592
OECD limitations in moving	7.5	5,819	560	24	21	23	44
OECD limitations in eating	4.0	5,118	653	2	9	29	38
Scores on SF-12 (age ≥ 12 years)							
The lowest score on mental health scales	9.6	2,732	270	-23	-9	-41	-27
A low score on mental health scales	19.1	2,461	102	-9	0	12	17
Limitations in daily activities (ADL) (age ≥ 55 years)							
At least one bad score on ADL scales	3.6	7,227	655	17	15	-23	-15
Presence of disease or disorder (age ≥ 12 years)							
Diabetes mellitus	5.5	5,191	-424	5	6	111	112
Stroke, brain infarction (ever)	2.5	5,274	-32	3	5	15	64
Myocardial infarction or other serious heart disease (ever)	3.1	6,425	545	37	29	29	20
Some type of cancer (ever)	7.3	4,895	408	4	2	-24	-20
Migraine or serious headaches regularly	15.4	1,925	10	-16	3	-17	-4
Hypertension	17.0	3,388	201	-4	1	10	22
Vascular construction (in stomach or legs)	2.1	5,178	325	28	24	2	5
Asthma, chronic bronchitis, lung emphysema	8.2	3,961	186	4	10	-40	-27
Psoriasis	2.9	2,294	165	-23	-12	-30	-7
Chronic eczema	4.5	2,510	-107	-6	6	102	105
Dizziness with falling down	3.2	3,568	135	-1	2	-26	1
Urine incontinence	5.7	4,310	503	5	12	14	31
Arthrosis of hips or knees	14.8	3,468	183	2	4	-11	-4
Rheumatoid arthritis	5.5	3,750	193	-22	12	-46	-4
Serious /persistent problems of neck or shoulder	10.5	3,107	253	-13	4	-10	1
Serious/persistent problems of hand, wrist or elbow	6.3	3,373	315	-9	5	-34	-17
Other long-term disease or disorder	11.5	4,335	460	-7	-1	80	91

Table A.6.4: (continued)

Groups	Size, in %	Average observed expenses, in €'s	Average residual expenses in year 2011, in €'s				
			Model 0 [b]	(Model 1-4 minus Model 0)			
				Model 1 [c]	Model 2 [d]	Model 3 [e]	Model 4 [f]
Co-morbidity (age ≥ 12 years)							
Two self-reported diseases or (chronic) disorder	6.5	2,307	-5	-13	-9	5	16
Health care utilization (all respondents)							
Glasses or contact lenses	37.1	2,275	-64	1	3	-12	-8
Hearing-aid	3.4	4,760	623	-6	5	-33	-32
Complete dentures	10.5	4,456	168	-1	4	-27	-14
Contact with home nurse (cure) in the past year	0.8	8,865	1,169	9	10	64	21
Home help assistance in the past year	1.5	4,247	536	-17	12	-13	-10

Footnotes Table A.6.4:

a. Residual expenses were calculated as: observed expenses minus predicted expenses of the Dutch RE-model of 2013. Expenses were annualized and weighted by the enrollment period. Residual expenses were calibrated: individuals' residual expenses per RE-model were raised by a factor equaling average RE-predicted expenses in the survey sample divided by average observed expenses in the survey sample. After this calibration, average residual expenses per RE-model were zero in the survey sample, just as is the case in the population. For Model 0, average residual expenses per group are presented. For Models 1 to 4, the difference in average residual expenses per group between these models and Model 0 are presented.

b. The statistics indicate the difference in average residual expenses for Model 1 to 4, respectively to Model 0. A negative value imply that average residual expenses of this model for this group are reduced and a positive value imply that average residual expenses are increased, compared to Model 0.

c. Model 0: the Dutch RE-model of 2013.

d. Model 1: Model 0 + a dummy variable for the persistently under-compensated group according to definition type 1 (i.e. positive residual expenses in each year over three consecutive years).

e. Model 2: Model 0 + a dummy variable for the persistently under-compensated group according to definition type 2 (i.e. top 50% of positive residual expenses in each year over three consecutive years).

f. Model 3: Model 0 + interaction terms for the persistently under-compensated group according to definition type 1 (i.e. positive residual expenses in each year over three consecutive years).

g. Model 4: Model 0 + interaction terms for the persistently under-compensated group according to definition type 2 (i.e. top 50% of positive residual expenses in each year over three consecutive years).

Chapter 7

Conclusions and Discussion

This chapter summarizes the main conclusions of the preceding chapters by answering the research questions formulated in the introduction, which leads to answering the central question of this thesis: *"How to evaluate the predictive performance of risk equalization models and to what extent can the predictive performance of a morbidity-based risk equalization model be improved by three new methods?"*. Next, the findings of this thesis are discussed. Finally, several recommendations and directions for further research are provided.

§ 7.1 HOW TO EVALUATE THE PREDICTIVE PERFORMANCE OF RISK EQUALIZATION MODELS?

§ 7.1.1 Which measures-of-fit have been used for evaluating risk equalization models and how have these measures been applied?

In order to answer the research question addressed in *Chapter 2*, a systematic literature review was conducted. This review resulted into a taxonomy of measures-of-fit for RE-models, together with a critical assessment of their properties and the analytic method of applying these measures.

In total, 71 different measures-of-fit have been used for evaluating RE-models since 2000. The taxonomy clusters these 71 measures into 30 measures by aggregating across four important variations in analytic method and then, these 30 measures into three categories based on the treatment of the prediction error. These three categories are: (i), measures based on squared errors: e.g. the R-squared (R^2) or the Mean Squared Prediction Error (MSPE); (ii), measures based on untransformed errors: e.g. the Mean Prediction Error (MPE) or the Predictive Ratio (PR); and (iii), measures based on absolute errors: e.g. the Mean Absolute Prediction Error (MAPE) or the Cumming's Prediction Measure (CPM). Four important variations in applying these measures are: the level of analysis, the type of sample used for prediction and model evaluation, the reference point against which predicted expenses are compared, and standardization. First, the level of analysis is concerned with whether the measures are applied at the individual level (i.e. the full sample), group level, or both. Almost all studies have applied measures at the individual level, with many of them applying measures at the group level as well. Second, most studies have conducted the evaluation on a validation sample (i.e. another sample than the one used for model estimation) in order to prevent overfitting, while several studies also have used the same sample for model evaluation as used for model estimation. Third, the conventional method for calculating the prediction error upon which the measures-of-fit are based is to use observed expenses as the reference point. Only a few studies have applied measures with a reference point other than observed expenses, namely normative expenses or predicted expenses calculated from a model with a broader set of risk adjusters than the one that is evaluated. Fourth, measures-of-fit can be standardized or unstandardized. Examples of standardized measures

are the R^2, PR, and CPM. Examples of unstandardized measures are the MSE, MPE, and MAPE. The value of these unstandardized measures change when expenses are expressed in another scale of measurement, while the value of standardized measures remains the same. Standardization is not a distinctive measurement property, because unstandardized measures can be standardized.

In sum, this review shows that there is no single measure-of-fit best across all situations. This is because the choice of applying a specific measure depends on preferences about the treatment of the prediction error and other measurement properties, together with the analytic method of applying the measure. However, if the purpose of evaluating RE-models is to estimate the extent to which the model achieves its policy goal – i.e. mitigating financial incentives for risk selection –, there is only one appropriate evaluation method, which is examining the prediction error for *selective groups* in the population that are as homogenous as possible.

§ 7.1.2 How to estimate the potential selection profits for multiple groups simultaneously under a risk equalization model?

In order to answer the research question dealt with in *Chapter 3*, different methods to estimate the potential selection profits for multiple groups simultaneously were investigated. As a starting point, a set of selective overlapping groups in the population were defined. This set of pre-defined groups was used to develop and apply three methods to estimate the potential selection profits for multiple groups simultaneously. These three methods define in different ways mutually exclusive groups from the same set of pre-defined groups. Aggregating residual expenses over all individuals that are assigned to the created mutually exclusive groups provide an estimate of the potential selection profits for multiple groups simultaneously under an RE-model. In addition, the three methods were compared to a relatively simple one, as used in previous studies, by aggregating average residual expenses for overlapping groups.

A first important finding is that the three methods based on mutually exclusive groups yield different estimates of the potential selection profits in monetary terms under each RE-model, but they lead to the same conclusion about which model yields the largest *overall* reduction in the potential selection profits. A second important finding is that estimating the potential selection profits by aggregating average residual expenses for overlapping groups does *not* lead to another conclusion about which model yields the largest *overall* reduction in the potential selection profits. To conclude, in this particular empirical application it did not matter which evaluation method was used to estimate the potential selection profits for multiple groups simultaneously – even usage of overlapping groups may satisfy – if the purpose is choosing among alternative RE-models (this does *not* hold if the purpose is interpreting how large the overall potential selection profits in monetary terms are under an RE-model).

§ 7.1.3 Conclusions

Using the findings from Chapters 2 and 3, an answer to the first part of the central question: *"How to evaluate the predictive performance of risk equalization models?"* can be provided by formulating several general principles. These principles as summarized in Table 7.1 may assist researchers and policymakers by performing an empirical evaluation and interpreting

Table 7.1: General principles for evaluating the predictive performance of risk equalization (RE) models

Principles regarding the design of the evaluation method and the measure(s)-of-fit:	
#1	Evaluations on selective groups are preferred over evaluations on individuals, random groups, or insurers' portfolios, because such group-level evaluations can indicate financial incentives for risk selection.
#2	Evaluating the model on multiple groups separately may be useful to provide a broad picture of model's predictive performance.
#3	When RE-models provide conflicting results because they perform differently on different groups, it may be helpful to evaluate the statistical fit on multiple groups simultaneously.
#4	For group-level evaluations, the selective groups should *not* be identical to the risk adjusters (in the form of dummy variables) in the RE-model [a].
#5	For group-level evaluations, the groups should be as homogeneous as possible and should not be too small in order to reduce the influence of random variation.
#6	The choice of the measure-of-fit depends on preferences about analytic method and the treatment of the prediction error, together with other measurement properties, such as whether the upper- or lower-bound of this measure for RE-models is known and the interpretation of the results of this measure in the specific context of RE.
#7	Measures based on untransformed errors, like the MPE or PR, are only meaningful when they are applied on selective groups [a].
#8	Since each measure has its pros and cons, it is useful to apply multiple measures at the same level of analysis.
Principles regarding the interpretation of the results of empirical evaluations:	
#9	One should be cautious with comparing the results across studies, when there are differences in datasets, settings, and/or methods.
#10	RE-models cannot, and do not have to, predict expenses adequately for each individual in the population, but they should predict expenses adequately for selective groups.
#11	The results of evaluations using measures with observed expenses as the reference point cannot, and do not have to, achieve the theoretical upper- or lower-bound of the measures, because observed expenses include a random component (i.e. unpredictability) and cost variation for which the regulator may not want to compensate.
#12	To interpret the results with the aim to examine the existence of under- and over-compensations for selective groups and hence, the presence of financial incentives for risk selection, a representative sample is *not* necessarily required.
#13	Average under- and over-compensations for groups do not have to equal zero, because of transaction costs and uncertainties around these estimates. Average under- and over-compensations for groups close to zero may suffice to prevent risk selection.
#14	For group-level evaluations, an under-compensation for one selective group implies an over-compensation for the complementary group [a].
#15	One should be cautious with interpreting the results for groups that are identical to the risk adjusters (in the form of dummy variables) in the RE-model [a].

Footnotes Table 7.1:

a. This principle only holds if the RE-model is estimated by Ordinary Least Squares (OLS), because in this case average predicted expenses equal average observed expenses for each group explicitly included in the model in the form of a dummy variable. This principle does *not* necessarily hold if other statistical specifications than OLS are used, such as two-part models, generalized linear models, or models based on (log-)transformed data. In this other case, it may be interesting to investigate how well the model predicts expenses for those groups explicitly included in the RE-model because they are selective groups that are of interest to the regulator.

the results of these evaluations. It is worth noting, however, that Table 7.1 is *not* exhaustive. This table presents those principles of which I think are crucial to evaluate the predictive performance for indicating to what extent an RE-model reduces financial incentives for risk selection and those that may take away misconceptions in the literature about some measures-of-fit in combination with their applications. Section 7.3 will reflect on these principles.

§ 7.2 TO WHAT EXTENT CAN THE PREDICTIVE PERFORMANCE OF A MORBIDITY-BASED RISK EQUALIZATION MODEL BE IMPROVED BY THREE NEW METHODS?

§ 7.2.1 New method 1: including risk adjusters based on cost and/or diagnostic information from multiple prior years

The findings in *Chapter 4* showed that the predictive performance of the Dutch RE-model of *2012* can be improved by including risk adjusters based on cost and/or diagnostic information from three prior years. The performance of this RE-model at the individual level in terms of the R^2 could be improved by approximately 8 percentage points and the MAPE could be reduced by approximately € 125 by including additional risk adjusters based on multiple-year cost and diagnostic information (benchmark: R^2 is 28.54%, MAPE is € 1,475). For RE-models that do not already include a risk adjuster for 'multiple-year high costs' (i.e. the MHC-groups as used in the Dutch RE-model of 2012), e.g. a demographic model or the Dutch model of 2011, there is even a larger potential for improving predictive performance by using multiple-year cost and diagnostic information. Besides an improved statistical fit at the individual level, the additional risk adjusters produce better predictions of expenses for several selective groups of interest. The predictions for some groups may even improve to such an extent that the under-compensations for these groups are no longer statistically significantly different from zero. This is the case, for example, for individuals who reported OECD limitations in moving, who had a low score on one of the SF-12 health scales, or who had limitations in daily activities.

However, the findings also showed that extending the Dutch model of 2012 with risk adjusters based on cost and/or diagnostic information from multiple prior years still led to (statistically significant) under-compensations for certain selective groups in the population, such as individuals who reported a poor general health status, individuals with a myocardial infarction or other serious heart disease, or individuals with at least one self-reported long-term disease. These findings lead to the conclusion that the new risk adjusters investigated here can reduce financial incentives for risk selection but cannot remove them completely, given the Dutch model of 2012.

§ 7.2.2 New method 2: including interaction terms between existing risk adjusters

In order to provide an answer to the research question dealt with in *Chapter 5*, regression tree models have been applied to identify all interaction terms between existing risk adjusters in the Dutch RE-model of *2014* that may have statistically significant predictive power in addition to the existing risk adjusters in this RE-model. The findings showed that the predictive performance of this RE-model can be improved by inclusion of interaction terms between existing risk adjusters. For this RE-model, the R^2 at the individual level can increase by 0.08 to 1.78 percentage points and the CPM by 0 to 0.44 percentage points, and the MAPE can reduce by € 0 to € 9, depending on the specification of the regression tree model (benchmark: R^2 is 25.56%, CPM is 24.98%, MAPE is € 1,569). Furthermore, the empirical results showed that better predictions of expenses for some selective groups in the population may be expected, e.g. for individuals who contacted a home nurse or those who used durable medical equipment in the previous year.

The findings, however, also showed that an RE-model that is extended with such interaction terms still (statistically significantly) under- and over-compensates insurers for several selective groups of interest. Examples of under-compensated groups are: individuals who reported a poor general health status, individuals who had a low score on the SF-12 physical health scales, or individuals with at least one self-reported long-term disease. These findings lead to the conclusion that interaction terms between existing risk adjusters can reduce financial incentives for risk selection but cannot eliminate them, given the Dutch RE-model of 2014. Another important finding of this study was that regression tree models are not robust with respect to the identification of the interaction terms. Consequently, it is not possible to draw conclusions about *which* specific interaction terms should be used in practice.

§ 7.2.3 New method 3: including a risk adjuster or interaction terms based on residual expenses from multiple prior years

The empirical results in *Chapter 6* first showed that a morbidity-based RE-model as the Dutch model of *2013* persistently under-compensates a selective group in the population over a three-year time period. Analyzing the expenses and risk characteristics of these persistently under-compensated individuals revealed that they generally have high above-average expenses in a year, have a high probability of being classified in a pharmacy-based cost group (PCG), diagnostic cost group (DCG), and/or a multiple-year high cost group (MHC-group), and often have at least one self-reported long-term diseases, with a high probability of having multiple diseases. The findings showed that extending the Dutch model of 2013 with a risk adjuster or interaction terms based on prior years' residual expenses increases the statistical fit at the individual level and group level. In this empirical analysis, an additional risk adjuster increases the R^2 at the individual level by 0.05 or 0.10 percentage points and the CPM by 0.08 or 0.47 percentage points, and reduces the MAPE

by € 1 or € 9, depending on the exact definition of the persistently under-compensated group (benchmark: R^2 is 24.22%, CPM is 24.88%, MAPE is € 1,570). Including interaction terms led to a larger improvement of the predictive performance of the benchmark model at the individual level: the R^2 increases by 1.51 or 1.64 percentage points, the CPM increases by 0.26 or 0.66 percentage points, and the MAPE reduces by € 5 or € 13, depending on the exact definition of the persistently under-compensated group. These findings show that the increase in percentage points in R^2 is smaller than the increase in percentage points in CPM when a risk adjuster is included, while the opposite holds when interaction terms are included, indicating that interaction terms are better able to predict expenses for some individuals in the (far) right-tail of the cost distribution than one additional risk adjuster. The empirical results also showed that inclusion of a risk adjuster or interaction terms may also lead to a better statistical fit for some selective groups in the population: e.g. individuals with at least one self-reported long-term disease, or those with serious back problems or back pain, or those who contacted a physiotherapist in the previous year. However, the under-compensations for these groups are still statistically significantly different from zero.

Furthermore, the findings showed that the RE-models with a risk adjuster or interaction terms defining the persistently under-compensated group may deteriorate the prediction of expenses for some other groups in the population. Examples of such groups are individuals who have limitations in hearing or seeing. These findings lead to the conclusion that an additional risk adjuster or interaction terms based on prior years' residual expenses can mitigate financial incentives for risk selection but cannot remove them completely, given the Dutch model of 2013.

§ 7.2.4 Conclusions

The findings from Chapters 4, 5, and 6 can be used to provide an answer to the second part of the central question: *"To what extent can the predictive performance of a morbidity-based risk equalization model be improved by three new methods?"*. All three methods are based on information that is, in most situations, already available in the administrative files of (Dutch) insurers; for instance, information on observed expenses, existing risk adjusters, and residual expenses. Consequently, these methods could be implemented with relatively low administrative costs as there are no additional costs for data collection. In some situations, however, diagnostic information may not be routinely available and hence, implementation of a method based on this information requires data collection.

Chapters 4, 5, and 6 lead to the conclusion that the predictive performance of a sophisticated morbidity-based RE-model can be improved by adding risk adjusters based on cost and/or diagnostic information from multiple prior years (*given* the Dutch model of 2012), or interaction terms between risk adjusters that are already included in the RE-model (*given* the Dutch model of 2014), or a risk adjuster or interaction terms based on residual expenses from multiple prior years (*given* the Dutch model of 2013). The exact extent to which the

predictive performance of the benchmark model can be improved at the individual and group level by each of these methods are presented in previous paragraphs (see § 7.2.1 – § 7.2.3). Hence, each of three methods investigated here can reduce financial incentives for risk selection, in particular for those groups that are explicitly identified by these additional risk adjusters or interaction terms. For these groups, financial incentives for risk selection are removed completely because for individuals in these groups average predicted expenses equal average observed expenses, if this model is estimated by OLS and the new risk adjusters or interaction terms are defined in the form of dummy variables. Since the dataset and benchmark model used differ across Chapters 4, 5 and 6, it is *not* possible to draw definite conclusions about which of the three methods has the largest predictive power.

The preceding chapters, however, also demonstrate that financial incentives for risk selection cannot be eliminated by each of the three methods alone, because each of them provides (statistically significant) under- and over-compensations for some selective groups in the population. This leads to the conclusion that implementing each of these new methods alone is *not* enough to achieve an RE-model that adequately compensates insurers for the expected expenses for each selective group that may be of interest to the regulator.

§ 7.3 DISCUSSION ON EVALUATING RISK EQUALIZATION MODELS

The first part of this thesis has formulated several general principles that can be used to evaluate the predictive performance of RE-models and to interpret the results of empirical evaluations. These principles, however, are *not* a blue-print for how to evaluate RE-models and interpret the results of empirical evaluations across all situations. This is because in each particular situation it is required to incorporate: (i), the specific study setting, e.g. the population and the year(s) that are analyzed; (ii), practical difficulties, specifically regarding the availability and quality of data; and (iii), the definition of the risk adjusters in the RE-models that are evaluated. Nonetheless, the systematic literature review in Chapter 2 of this thesis revealed a number of misconceptions, which illustrate the need for general principles that can be used for conducting empirical evaluations and interpreting the results of such evaluations. Misconceptions refer to situations where measures have not been applied appropriately or situations where the results of these measures have not been interpreted appropriately. The next paragraph provides some evident examples of misconceptions. After this, the general principles that can be used for choosing the design of the evaluation method and the measure(s)-of-fit will be discussed (see § 7.3.2), followed by those that can be used for interpreting the results of empirical evaluations (see § 7.3.3). Finally, this section discusses that deciding on the design of RE-models involves incorporating more evaluation criteria than solely the predictive performance.

§ 7.3.1 Misconceptions

A first evident misconception, which requires specific attention, relates to the definition of the groups on which RE-models have been evaluated. Evaluating model's performance on groups that are identical to those included in the model under study is not meaningful if OLS is used. As previously mentioned, any OLS-model perfectly predicts expenses for all groups that are explicitly defined by the risk adjusters in the form of dummy variables when this model is estimated on the total dataset. Note that if a validation sample is used, average predicted expenses may be very close to average observed expenses for all groups explicitly included in the model. This evaluation method is in particular misleading when the evaluation-groups are based on information that is used by *one* of the evaluated RE-models, because that model will automatically perform (much) better on these groups than the other models, which eventually may influence the conclusion about which model to use. Another example is the calculation of PRs (or any other measure-of-fit) for groups based on intervals of *predicted* expenses. The PR for these intervals will approach unity, regardless of the quality of the RE-model, if OLS is used. This is because OLS results in a linear relationship between predicted expenses and observed expenses with a slope coefficient of one. Another example closely related to the previous one is plotting observed expenses (Y-axis) against predicted expenses (X-axis), whereby the X-axis is divided in intervals based on the predicted expenses of *one* of the evaluated RE-models. The RE-model that is used to define these intervals will outperform the others, because the groups are based on this model. These evaluation methods may lead to misinterpretation of models' true performance in terms of the statistical fit for specific groups in the population that may be of interest to the regulator.

A second evident misconception is the comparison of R^2-values (or any other measure-of-fit) across studies, when there are differences in datasets, settings, and methods. Differences in the R^2-values can be misinterpreted as differences in the predictive power of the set of risk adjusters included in the models under study, while they may very well be due to other factors.

The aforementioned examples show that the evaluation methods and the properties of measures in combination with how these measures are applied have not always been well-understood. This illustrates the relevance of the general principles formulated in Table 7.1, specifically regarding the definition of the evaluation-groups (see *principles #4, #5,* and *#15*) and the comparison of the results of evaluations across studies (see *principle #9*).

§ 7.3.2 Principles regarding the design of the evaluation method and measure(s)-of-fit

In Table 7.1, principles #1 to #8 are concerned with choosing the design of the evaluation method and the measures-of-fit. These principles are not only useful to those who conduct evaluations but they are also useful to those who interpret the results of evaluations for decision-making. In order to choose among alternative RE-models it is required to have

a good understanding of how the models have been evaluated in order to judge the appropriateness of this evaluation method for its study-purpose and hence, the importance of the results that are found.

Principles #1 – #3: group-level evaluations

Chapter 2 in this thesis has concluded that there is only one adequate evaluation method if the purpose is to estimate financial incentives for risk selection, which is evaluating the statistical fit of the RE-model for *selective groups* in the population (see *principle #1*). Other evaluation methods, such as evaluations at the level of individuals, random groups, or insurers' portfolios are not adequate for this purpose. This is because the results of these methods can be the net effect of (significant) under- and over-compensations for selective (homogeneous) groups. For the results at the individual level, the upper- or lower-bound of measures-of-fit for RE-models is also generally unknown: e.g. the R^2 or CPM. The results at the portfolio level also depend on the accidental risk composition of insurers' portfolio. Consequently, the results of these types of evaluations may change if a selective group in the population switches from insurer, while the RE-model under study remains the same.

To provide a broad picture of the financial incentives for risk selection under an RE-model, it is useful to evaluate the predictive performance for *multiple* selective groups of interest (see *principle #2*). For each pre-defined group a separate result is obtained per model, which indicates how well this model predicts expenses for this group. Such evaluations can provide highly valuable information about the presence of financial incentives for risk selection under the RE-model(s) under study and how to further improve the statistical fit.

In some situations, however, it is helpful to apply one single measure reflecting model's predictive performance on multiple groups simultaneously as developed in this thesis (see *principle #3*). This is because RE-models can perform differently on different groups and hence, conflicting results may be obtained. Eventually, it may be difficult to decide which model to use. A single measure that summarizes the *overall* financial incentives for risk selection for multiple groups may help decision-making.

The empirical evidence provided by this thesis, however, may be too thin to draw strong conclusions about which specific method for estimating the potential selection profits for multiple groups simultaneously is best across all situations. Alternative sets of pre-defined groups and alternative RE-models were not examined. Consequently, it is preferred to apply the evaluation methods based on mutually exclusive groups examined here all together (including the simple method of aggregating residual expenses on overlapping groups) in order to investigate whether it matters which method is used to choose among alternative RE-models, given the set of pre-defined groups and the models that are evaluated in this particular situation. If more empirical evidence is obtained across different settings, it may be concluded with more confidence whether one evaluation method outperforms others and if so, which method can be used best.

Furthermore, although a single measure reflecting model's performance for multiple groups simultaneously can be helpful for decision-making, it may still be meaningful to apply additional measures to examine the results on separate groups. Such detailed evaluations can indicate which underlying (sub)groups are responsible for the under- and over-compensations for multiple groups together, providing valuable information for potential improvements of the RE-model under study.

Principle #4 – #5: definition of the selective groups of interest

A crucial element of conducting group-level evaluations is the definition of the groups. As noted earlier, some misconceptions exist about the definition of the evaluation-groups and therefore, the next two principles specifically focus on this element.

A first important principle for defining selective evaluation-groups is that they should *not* be identical to the risk adjusters in the form of dummy variables in the RE-models that are evaluated, if these models are estimated by OLS (see *principle #4*). So far, OLS is the conventional estimation technique in the field of RE. It is worth noting that this is *not* the case when another statistical specification than OLS is used, e.g. two-parts models, generalized linear models, or 'constrained regression', which is a recently developed technique by van Kleef and colleagues (van Kleef et al., 2015). For model specifications other than OLS, it may be helpful to evaluate the statistical fit on the same groups as included in the model under study because this model may under- or over-compensate insurers for these selective groups that are of interest to the regulator.

A second important principle for defining the selective groups of interest is that they should be as homogeneous as possible (see *principle #5*). This is because otherwise the under- and over-compensations for the groups can be the net effect of (significant) under- and over-compensations for homogeneous subgroups. Furthermore, the groups should not be too small in order to reduce the chance of random fluctuations.

The aforementioned two principles may impose restrictions on the type of information that can be used for evaluating RE-models. In general, evaluation-groups can be derived from any type of information that may be used (indirectly) by an insurer or consumer for exploiting risk selection, as long as the two previously mentioned principles are taken into account. For example, this could be information from insurers' administrative files or health surveys.

Given the availability of information to define selective groups of interest, it is important to decide for *which* groups the statistical fit is evaluated. From a regulator's perspective, this choice may be guided by the question which groups can be subject to risk selection. From an insurer's perspective, however, any selective group with an under- or over-compensation may be of interest, regardless whether the regulator wants compensation for this group or not (i.e. the S-type and N-type risk factors, respectively). Consequently, even when an RE-model adequately predicts expenses for any selective group that is of interest to the regulator,

there still may be under- and over-compensations for other groups that are of interest to the insurer because they may determine healthcare expenses, e.g. groups based on lifestyle factors such as smoking. Even though the regulator may not want compensation for these other groups, it may be of interest to monitor the under- and over-compensations for these groups because they may be vulnerable to risk selection. Furthermore, they may correlate to groups for which the regulator wants compensation: e.g. smoking may determine illness. If these other groups are completely determined by N-type risk factors, it may be useful to relax premium rate restrictions for these groups, e.g., by allowing a premium bandwidth as used in the U.S., rather than creating financial incentives for risk selection on these groups, which may jeopardize solidarity, efficiency, and quality of care. Note, however, that it may be (very) difficult to disentangle S-type from N-type cost variation in practice (Schokkaert & van de Voorde, 2004, 2006, 2009). If the expenses of groups are partly determined by N-type risk factors, even though it is unknown to what extent, the average under- or over-compensation for these groups do not have to equal zero in practice, as will be discussed below.

Principles #6 – #8: choice of the measure(s)-of-fit

Chapter 2 has concluded that there is no single measure best across all situations, because the choice of a specific measure depends on preferences about the measure's properties and the analytic method (see *principle #6*). It is beyond the scope of this thesis to fully review all measures' properties in relation to the method of applying these measures. Here, specific attention will be paid to a distinctive measurement property and to some properties that are important to take into account in combination with the analytic method for the purpose of indicating financial incentives for risk selection.

A distinctive measurement property is the treatment of the prediction error and therefore, the choice of a specific measure requires an explicit judgment about this property. If it is preferred to weigh prediction errors differently, measures based on squared errors like the R^2 or MSPE may be appropriate. For these measures, an improvement of the prediction of expenses for individuals with large errors (i.e. outlier observations) will lead to a relatively large increase in the value of these measures because large errors are more heavily weighted than small errors. If, however, it is not preferred to weigh errors differently, measures based on untransformed errors like the MPE or PR, or measures based on absolute errors like the CPM or MAPE may be appropriate. These measures are less sensitive to outlier observations.

In addition to the treatment of the prediction error, other measurement properties may play an important role in deciding which measure(s) to use because they may interrelate to the application of the measure(s). An important property of measures based on untransformed errors is that negative errors may cancel out positive errors. This property requires specific attention because measures based on untransformed errors have been applied regularly to evaluate the statistical fit for groups. The extent to which this happens depends on the heterogeneity in terms of residual expenses for the population or group that is ana-

lyzed (see *principle #5*). An evident example is examining the MPE or PR for an RE-model that is estimated by OLS at the individual level (i.e. the estimation sample as a whole): the MPE and PR always equal zero or one, respectively, no matter the predictive performance of this model. Therefore, measures based on untransformed errors are only meaningful at the level of selective groups (random groups or random insurers' portfolios will also not suffice) (see *principle #7*). For the same reason it is important that the pre-defined selective groups are as homogeneous as possible (see *principle #5*). With measures based on absolute errors, however, it is not possible that positive errors cancel out negative errors and so, these measures can be applied at both the level of individuals and groups. However, for these measures – e.g. the MAPE and CPM – the lower- or upper-bound for RE-models in practice is generally unknown, making it difficult to interpret the results to indicate to what extent the predictive performance can be improved further. In contrast to the MAPE and CPM, however, the MPE at the level of groups has an intuitive interpretation because the average under- and over-compensations are presented in monetary terms and the lower-bound of this measure for RE-models is known, namely zero.

The discussion of measures' properties demonstrate that each measure has its pros and cons and for this reason, it is useful to apply multiple measures at the same level of analysis for evaluating alternative RE-models (see *principle #8*). For a comparative model evaluation at the *individual level*, it may be useful to apply the R^2 in combination with a CPM and/or MAPE, despite these measures cannot indicate financial incentives for risk selection. If there is a relatively large increase in R^2 but the CPM and/or MAPE only marginally changes, it may be an indication that the prediction of expenses for some groups with large residual expenses has improved while the prediction for others has deteriorated (Chapters 5 and 6 are examples of this phenomenon). This is because the CPM and MAPE weigh errors equally, while the R^2 weighs large errors more heavily. Consequently, the improved prediction of expenses for some individuals reflected in the R^2 should have led to a deteriorating of the prediction for others because otherwise the CPM or MAPE should also have changed. If, however, the MAPE and/or CPM at the individual level also change relatively much it may indicate that the prediction of expenses for groups is improved (see Chapter 4 for an example). For a comparative model evaluation at the *group level*, it may be useful to combine measures if the pre-defined groups are heterogeneous. For example, the MPE and MAPE can show whether the observed under- and over-compensations for the heterogeneous groups may be the net effect of under- and over-compensations for (homogeneous) subgroups. If the MPE equals the MAPE for the same group, then this group is perfectly homogeneous in terms of residual expenses.

§ 7.3.3 Principles regarding the interpretation of the results of empirical evaluations

In Table 7.1, principles #9 to #15 are concerned with interpreting the results of empirical evaluations. Generally, empirical evaluations consist of a comparison of the performance of

alternative RE-models. The results of such evaluations can be interpreted for the purpose of choosing which of the evaluated models has the highest predictive performance and hence, should be used. The interpretation of the results, however, largely depends on the specific evaluation method that has been used, including the measure(s)-of-fit that has been applied. Below first a principle is discussed that holds for any type of evaluation, followed by principles for interpreting the results of evaluations at the level of individuals and those of evaluations at the level of groups.

Principle #9: any type of evaluation

A principle that applies to any type of evaluation is that one should be cautious with interpreting the results across studies, when there are differences in datasets, settings and/or methods (see *principle #9*). As noted earlier, comparing the results across studies may lead to misinterpretation of models' true predictive performance. In order to make an appropriate comparison, alternative RE-models should be evaluated under the same conditions, such as the study setting and the dataset.

Principles #10 – #11: evaluations at the individual level

If a comparative model evaluation is conducted at the level of individuals, the results can only be interpreted to indicate which of the evaluated RE-models performs best in terms of the statistical fit for all individuals in the population; and *not* to indicate the financial incentives for risk selection. Generally, a higher predictive performance at the individual level may lead to reduced financial incentives for risk selection for groups in the population. To interpret the results of evaluations at the individual level, it is important to take into account that an RE-model cannot, and does not have to, predict expenses adequately for each individual in the population (see *principle #10*). Hence, an RE-model need not achieve the theoretical upper- or lower-bound of the measures that are applied, if they use observed expenses as the reference point for calculating the prediction error because observed expenses include a random component (i.e. unpredictability) (see *principle #11*). For example, an RE-model cannot achieve an R^2 at the individual level of one. These principles imply that there will always be under- and over-compensations for individuals, even under an adequate RE-model. These under- and over-compensations for individuals, however, are not of particular interest because an RE-model should adequately predict expenses for selective groups in the population.

Principles #12 – #15: evaluations at the group level

If a comparative model evaluation is conducted at the level of selective groups, the results can be interpreted for the purpose of indicating which of the evaluated RE-models yields the largest reduction in the financial incentives for risk selection and hence, is the preferred model to be used. One should be cautious with interpreting the results of group-level

evaluations for quantifying how large the incentives are in absolute money terms in the population. For such an interpretation of the results it is necessary that the sample used for evaluating model's performance is representative for the population. A representative sample, however, is *not* required to indicate whether there are financial incentives for risk selection under the RE-model under study and if improvement of this model is required (see *principle #12*). This is because insurers may have financial incentives to select groups that are (statistically significantly) under- or over-compensated, no matter they have a representative sample of the total group in the population.

An RE-model that yields the largest (overall) reduction in the average under- or over-compensations for selective groups that are of interest to the regulator can be considered the preferred model to be used. An RE-model can be considered to be adequate when average under- and over-compensations for selective groups of interest are close to zero (see *principles #11* and *#13*). These average under- and over-compensations do not have to equal zero because of: (i) transaction costs for engaging in risk selection (van Barneveld et al., 2000); (ii), statistical uncertainties around the net benefits of risk selection, e.g. high-risk individuals may become low-risk individuals and vice versa (Beck & Zweifel, 1998; Beck et al., 2010; Breyer et al., 2012; van Barneveld et al., 2000; van Kleef et al., 2013a; Welch, 1985); and (iii), the under- and over-compensations may be partly due to N-type cost variation, if observed expenses are used as the reference point for calculating these under- and over-compensations (random errors are expected to cancel out at the group level). In practice, however, it may be difficult to disentangle cost variation related to S-type and N-type risk factors and hence, it is unknown to what extent under- and over-compensations may be due to N-type risk factors (Schokkaert & van de Voorde, 2004, 2006, 2009). Furthermore, it is important to realize that in case of OLS an under-compensation for one group implies an over-compensation for the complementary group (see *principle #14*). In addition, one should be cautious with interpreting the results on groups that are identical to the risk adjusters in the RE-model, if this model is estimated by OLS and the risk adjusters are included as dummy variables (see *principle #15*). Not taking into account this principle may lead to misinterpretation of model's true predictive performance for selective groups of interest (see § 7.3.1).

§ 7.3.4 Other evaluation criteria in addition to the predictive performance

The first part of this thesis has focused on evaluating the predictive performance, which has received most attention in the literature over the past decades of all evaluation criteria but it is only one criterion. When deciding on the design of an RE-model, it is required to incorporate several criteria, as listed in Table 1.1 (see Chapter 1, page 22). Consequently, an RE-model with a higher predictive performance is not necessarily preferred over an RE-model with a lower predictive performance. Eventually, the same results of evaluations

may lead to different decisions about which model to use across policymakers in different countries because of different value judgments regarding the evaluation criteria.

§ 7.4 DISCUSSION ON IMPROVING RISK EQUALIZATION MODELS BY THE THREE NEW METHODS

The second part of this thesis has concluded that the three new methods can improve the predictive performance of the benchmark model under study. It is, however, not possible to draw conclusions about which of the new methods leads to the largest reduction in the financial incentives for risk selection and hence, is the preferred one to be used. This is because the administrative data and the benchmark model are not consistent across the studies. Consequently, differences in the results are not only due to differences in the predictive power of the risk adjuster or interaction terms but can also be due to differences in the dataset and analytic method (see *principle #9*). This section will provide some reflections on the results of the three new methods and discuss the generalizability.

§ 7.4.1 Reflections on the results of the three new methods

It was expected that each of the three methods investigated here would lead to (substantial) improvements of the statistical fit for several selective groups in the population. However, these improvements are not overwhelming, specifically regarding the interaction terms between existing risk adjusters (see Chapter 5) and the risk adjusters or interaction terms based on prior years' residual expenses (see Chapter 6). For each method, the benchmark model is extended with a large array of additional risk adjusters or interaction terms – sometimes the extended models includes more than double the number of original risk adjusters – and advanced methods on large datasets were needed to define these risk adjusters or interaction terms. This lead to the conclusion that it requires a relatively large amount of effort to accomplish better predictions for some selective groups in the population, given an RE-model that is already quite sophisticated with several morbidity-based risk adjusters. The following three potential reasons for these results can be formulated.

A first reason may be that improving the prediction of expenses for some groups leads to a deterioration for others. Consequently, the new method does not lead to a significant *overall* improvement in model's performance, for example as indicated by a marginal increase in the CPM or MAPE at the individual level. The results in Chapter 6 provide a clear example of this phenomenon.

A second reason may be that the new risk adjusters or interaction terms may not be specific enough to identify those individuals for whom the benchmark model does not predict expenses adequately (for example, the interaction terms between existing risk adjusters) and/or the new identified groups may be too heterogeneous with respect to residual

expenses (for example, the risk adjusters or interaction terms based on prior years' residual expenses). Furthermore, some individuals for whom the benchmark model does not predict expenses adequately may not be classified in the new risk adjusters or interaction terms, which may partly be due to restrictions on classifications (see Table A.1 on page 255 and Chapters 4 to 6).

A third reason may be that it is very difficult to provide adequate cost predictions for some specific groups in the population, such as individuals with rare diseases, those who are in the final stage of life, or pregnant women. Mortality and pregnancy are both not predictable (from a regulator's point of view) but these groups are sure to have high above-average expenses that may not be adequately captured by the risk adjusters and/or interaction terms in the model under study. Each of the new methods investigated here may identify some of these individuals; however, they may not be specific enough to identify all individuals in these 'problematic' groups. Though the results in this thesis are based on the Dutch RE-model, they may be exemplary for other RE-models used around the world.

§ 7.4.2 Generalizability of the results

Definite conclusions about the extent to which the predictive performance of RE-models used in practice can be improved by each of the three methods examined here are not possible because: (i), the quality of benchmark model influences the results; (ii), it is required to incorporate other evaluation criteria in addition to the predictive performance, such as value judgments about appropriateness of incentives and fairness that determine which risk adjusters and/or interaction terms are used; and (iii), the feasibility-criterion, such as the availability and quality of data, imposes constraints on the type of risk adjusters and/or interaction terms used. Each of these factors will be discussed below. As a result of this, it is required to conduct the empirical analyses in each specific context in order to conclude to what extent the predictive performance of the benchmark model can be improved by new risk adjusters and/or interaction terms and whether it is worthwhile to implement them.

Quality of the benchmark model

Using another RE-model as the benchmark used throughout this thesis, which is the Dutch RE-model, may yield different results with respect to the predictive power of the new risk adjusters and interaction terms. It is expected that for less sophisticated RE-models the potential improvement may be larger than found here, holding other things equal. For an RE-model which is similar with respect to the type of risk adjusters as the Dutch RE-model – e.g. those used in Belgium, Germany, and the U.S. – it is unknown to what extent these models can be improved by using each of the new methods. This is because it depends on the exact definition of the existing risk adjusters, the predictive power of these risk adjusters, and the availability and quality of data.

Value judgments about appropriateness of incentives-criterion and fairness-criterion

As noted earlier, the decision to implement a certain risk adjuster or interaction term is not solely based on its predictive power but involves other evaluation criteria as well (see Table 1.1, Chapter 1, page 22). Regarding the appropriateness of incentives-criterion, the new risk adjusters and interaction terms may mitigate financial incentives for risk selection. However, the risk adjusters or interaction terms based on prior years' (residual) expenses may reduce incentives for efficiency. Such risk adjusters or interaction terms have been highly debated in the literature because they may reward an inefficient insurer and penalize an efficient insurer (e.g. Ash et al., 1989; Lamers & van Vliet, 1996; Lamers, 1997; van Vliet & van de Ven, 1992, 1993, van de Ven & Ellis, 2000). Consequently, implementing these risk adjusters or interaction terms involves a trade-off between risk selection and efficiency. If policymakers consider the financial incentives for risk selection to be large compared to the reduced incentives for efficiency, it is possible to impose constraints on classification in these risk adjusters or interaction terms; for instance, only identify individuals with (residual) expenses above a certain threshold. Such constraints may also reduce the possibilities for manipulation.

Besides prior years' (residual) expenses, diagnostic information based on prior healthcare utilization has also been debated because it may stimulate more utilization than strictly necessary and may reward inefficient insurers (e.g. Ash et al., 1989; Lamers & van Vliet, 1996; Lamers, 1997). To reduce these incentives to a large extent, diagnostic information is used in practice by constructing cost groups. These groups are based on a classification algorithm of similar diagnoses (Ash et al., 1989; Ellis & Ash, 1995; Lamers, 1997). In addition, risk adjusters based on diagnostic information from one prior year already have been used in practice and so, using diagnostic information from multiple prior years may also be appropriate to use. The same argument applies to interaction terms between existing risk adjusters.

The fairness-criterion requires value judgments about the S-type and N-type risk factors. In Chapters 4 through 6 of this thesis, this criterion has not been taken into account, implicitly assuming that the new risk adjusters and interaction terms only capture S-type cost variation. It is likely that the new risk adjusters and interaction terms, however, may also partly adjust for N-type cost variation, especially those risk adjusters or interaction terms based on prior years' (residual) expenses, although it is unknown to what extent.

Feasibility-criterion

Regarding the feasibility-criterion, the new methods are based on information that is, in most situations, already available in the administrative files of (Dutch) insurers. Since there are no additional costs for data collection, it may be attractive to implement them in order to reduce financial incentives for risk selection. In some situations, however, not all methods can be implemented because the required data are not routinely collected or the quality

of data may be insufficient; for example, diagnostic information for the total population. Consequently, availability and quality of data pose restrictions on the types of risk adjusters and/or interaction terms that can be implemented.

Besides availability and quality of data, it is crucial to consider the acceptability, validity, and reliability of the new risk adjusters and/or interaction terms. The risk adjusters or interaction terms based on prior years' (residual) expenses are expected to identify those individuals in the population with a higher healthcare need and define relatively homogeneous groups in terms of (residual) expenses (i.e. validity). Hence, using these risk adjusters or interaction terms may not undermine the acceptability among stakeholders due to apparently trivial classification algorithms. This, however, may not hold for the new interaction terms between existing risk adjusters. These interaction terms are based on a complex classification algorithm that is solely based on statistical power rather than clinical judgment. Usage of such an automatic algorithm may undermine the acceptability under different stakeholders. Further, the validity and reliability of these interaction terms may be questioned, because regression tree modelling as used for defining the interaction terms is not robust. Consequently, the acceptability, validity, and reliability may play an important role in addition to the predictive power for deciding whether to use interaction terms and if so, which interaction terms to use.

To conclude, implementing the new risk adjusters and/or interaction terms is a complex undertaking. Different evaluation criteria need to be incorporated that may be in conflict, which eventually may lead to (inevitable) trade-offs between risk selection and efficiency, given the practical limitations on finding risk adjusters and/or interaction terms.

§ 7.5 RECOMMENDATIONS

The previous sections have summarized and discussed the findings of this thesis, which leads to six recommendations. First, two recommendations focus on evaluating the predictive performance of RE-models. Then, three recommendations focus on reducing financial incentives for risk selection by improving the predictive performance of the RE-model used in practice and by implementing alternative strategies. A final recommendation focusses on an effective coordination between 'model evaluation' *and* 'model improvement' in practice, while this thesis has investigated them as separate topics.

§ 7.5.1 Evaluation method to indicate financial incentives for risk selection

Group-level evaluations should be a fundamental part of each empirical study

This thesis clearly shows that the policy goal of RE requires a more advanced evaluation method than the conventional methods that have been suggested in statistical theory for evaluating prediction models. In order to quantify the extent to which an RE-model mitigates financial incentives for risk selection, it is required to evaluate the predictive performance on selective groups of interest. Therefore, it is recommended to conduct a group-level evaluation in each empirical study. Researchers and policymakers should recognize the importance of conducting such evaluations. The general principles as formulated in this thesis may contribute to this by creating awareness of the (in)appropriateness of some evaluation methods and the application of some measures-of-fit for indicating financial incentives for risk selection. Eventually, group-level evaluations should be a fundamental part of each empirical study rather than being applied infrequently as have been done so far.

Invest in collecting the required data on a routinely basis to conduct group-level evaluations

To perform a group-level evaluation, it is required to have external information in order to define groups that are not identical to the risk adjusters in the evaluated RE-model(s), if this model is estimated by OLS. This evaluation method, however, may not be routinely applicable because of lack of the required data. Because it is of great policy relevance to know to what extent an RE-model reduces financial incentives for risk selection, it is recommended to invest in collecting information for performing group-level evaluations, e.g. by conducting health surveys. Ideally, the same type of information is collected routinely because then the financial incentives for risk selection for the same selective groups of interest under the RE-model as used in practice can be monitored over time.

§ 7.5.2 Mitigating financial incentives for risk selection by improving the risk equalization model used in practice and alternative strategies

A critical question for policymakers involved in RE is how to further improve the predictive performance of sophisticated morbidity-based RE-models. This thesis demonstrates that it requires a relatively large amount of effort to improve the prediction for some selective groups in the population, which raises the concern to what extent RE-models can be further refined, *without* making the RE-model unnecessary complex with apparently trivial classification algorithms and/or an extensive set of risk adjusters and/or interaction terms. For the upcoming years, this may be one of the biggest challenges in the field of RE. The following recommendations focus on different methods to mitigate financial incentives for risk selection.

Financial incentives for risk selection can be mitigated by each of the three methods

This thesis shows that there is room for improving the predictive performance of a morbidity-based RE-model by including additional risk adjusters and/or interaction terms based on information that is already available. Each of the three methods examined may lead to a better statistical fit for some selective groups of interest and hence, may mitigate financial incentives for risk selection. Out of the three methods, implementation of risk adjusters based on diagnostic information from multiple prior years may be a first attractive method if this information is routinely available. This is because this type of information from one prior year is already used in several countries and multiple-year diagnostic information may have predictive power as shown in this thesis. The other methods examined here may lead to some (inevitable) trade-offs, which require value judgments of policymakers about whether or not to implement them.

Financial incentives for risk selection cannot be eliminated by each of the three methods alone: further research is required

This thesis also shows that if policymakers decide to implement one of the methods examined here there may be still under- and over-compensations for specific groups of interest. This raises the question to what extent the predictive performance can be improved when all three methods are combined, especially in order to investigate to what extent they are supplements or complements. If all three methods together still provide under- and over-compensations for selective groups of interest, which is not unlikely, it may be concluded that using information in insurers' administrative files alone may be not sufficient. Hence, it may be necessary to find additional risk adjusters based on other data sources. Possibly, further improvements can be found in collecting specific information for groups of individuals with rare diseases, individuals in the final stage of life, and/or pregnant women, because these groups are sure to have high above-average expenses and may have high (persistent) under-compensations. Further research on improving the predictive performance of RE-models used in practice is required, see § 7.6 for suggestions.

Consider (temporary) alternative strategies to mitigate financial incentives for risk selection

As long as RE-models used in practice do not adequately predict expenses for selective groups of interest, an alternative effective strategy to mitigate financial incentives for risk selection on specific groups in the population is the implementation of (temporary) risk sharing arrangements. For example, insurers may bear less financial risk for the approximately 1% - 4% of the total population who exhibit large under-compensations persistently over time; for the remainder of the population insurers bear 100% financial risk. In this case, residual expenses for the 1% - 4% of the population are retrospectively equalized to a pre-defined extent. This strategy may not provide optimal incentives for efficiency, but

it completely removes financial incentives for risk selection targeted on this persistently under-compensated group, which may be vulnerable to risk selection. Risk sharing arrangements can be implemented temporarily, until adequate risk adjusters and/or interaction terms are developed.

§ 7.5.3 Model evaluation and model improvement together

An effective coordination between evaluating and improving the predictive performance of risk equalization models is crucial

Finally, an effective coordination between evaluating and improving the predictive performance may lead to breakthrough improvements in mitigating financial incentives for risk selection for particular groups in the population. Insurers may act upon the predictive performance of the RE-model used, specifically regarding groups with high (persistent) under- or over-compensations. Therefore, it is important to monitor the potential selection profits under the RE-model used and if necessary, improve the design of this model upon the results of this analysis. Improving model's performance may reduce the information surplus between the regulator and insurers and hence, may reduce possibilities to exploit extra information for engaging in risk selection. It is also important to monitor trends in risk selection actions, because this may provide valuable information on which selective groups should be monitored over time. Monitoring these groups may indicate to what extent they are persistently under- or over-compensated. Eventually, these evaluations may provide insights into whether and how to improve the design of the RE-model used, specifically regarding those groups that are vulnerable to risk selection. In sum, evaluation and improving the predictive performance should be effectively coordinated in a continuous process.

§ 7.6 DIRECTIONS FOR FURTHER RESEARCH

Given the importance of an adequate RE-model, further research is needed to investigate how the predictive performance of currently-used RE-models can be improved further. Future research should pay particular attention to the 'most-problematic' groups in the population for whom the RE-model as used in practice under-compensates insurers persistently over time. It is relevant to investigate the risk characteristics of these individuals in much more detail as done in this thesis; for instance, by using diagnostic information underlying the PCGs and DCGs in combination with information about the use of specific healthcare facilities and external data sources, such as mortality. In addition, it is relevant to investigate alternative strategies for these 'problematic' groups, such as risk sharing arrangements. This is because it is not unlikely that these groups consist of individuals for whom

it may be very difficult to predict expenses adequately; for instance, individuals with rare diseases, individuals in the final stage of life (i.e. mortality), and pregnant women. Further research can provide useful insights on potentially relevant methods to mitigate financial incentives for risk selection, through new risk adjusters and/or new interaction terms and/or risk sharing arrangements.

Furthermore, further research waits to investigate to what extent fully exploiting all available information in the administrative files of insurers together can lead to improving the predictive performance of the model used in practice, while taking into account the different evaluation criteria when developing these risk adjusters and/or interaction terms. For doing so, it is important to involve stakeholders during the process of developing these risk adjusters and/or interaction terms because this may prevent problems with the acceptability, validity, and reliability when implementing them. For example, it is crucial to involve medical experts early in the process of developing interaction terms, aiming to identify specific patient groups guided by clinical judgment rather than statistical power. In addition, further research waits to investigate which other data sources are available and how this information can be used for RE.

General Appendices

GENERAL APPENDIX 1: RISK EQUALIZATION IN THE NETHERLANDS

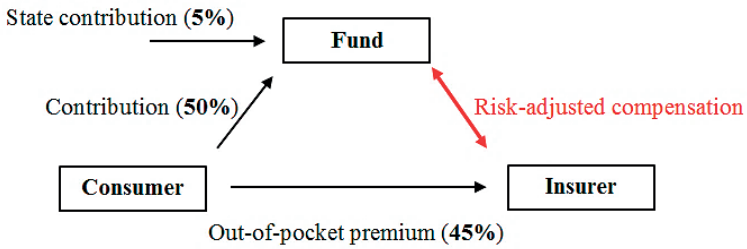
As noted, RE is used in several countries world-wide but the exact calculation of the risk-adjusted payments differs across countries, depending on the modality that is implemented and the design of the RE-model (van de Ven & Ellis, 2000). For example, the risk-adjusted payments can equal predicted expenses of an individual as calculated by the RE-model; e.g. this is the case in Israel, or it can equal predicted expenses minus the average national premium individuals pay their insurer out-of-pocket; e.g. this is the case in the Netherlands. Country-specific differences in the modality of the RE scheme, however, are not of particular interest for the objective of this thesis, since in all countries, regardless of modality, it is important that the RE-model adequately predicts individuals' expenses. The design of the RE-model can vary from relatively simple demographic models to more sophisticated morbidity-based models. This appendix briefly describes the organization of the RE scheme in the Netherlands and the design of the Dutch RE-model.

It is important to mention that the process of calculating the *actual* risk-adjusted payments is country-specific with many technical details and policy regulations and rules. These country-specific details are left out of the description in the next sections, as much as possible, because they make the description unnecessarily complex for the purpose of understanding the Dutch RE scheme and the design of the Dutch RE-model.

§ A.1.1 Organization of the RE scheme

Figure A.1 graphically depicts the organization of the RE scheme in the Netherlands. The RE Fund is filled with mandatory contributions via (income-related) tax revenues for all Dutch citizens of 18 years and older and a state contribution for children and adolescents. In total, insurers receive 50% of their revenues via risk-adjusted compensations from the RE Fund; the other 50% of their revenues are obtained via out-of-pocket premiums for individuals of 18 years and older. These out-of-pocket premiums are community-rated, implying that insurers must charge the same premium for the same insurance product, regardless of individuals' expected healthcare expenses. The risk-adjusted compensations are estimated by the Dutch RE-model as described in the next sections. Insurers receive more money from the RE Fund when they have an above-average proportion of high-cost individuals, such as elderly or chronically ill (i.e. a high-risk portfolio); inversely, insurers receive less money when they have a below-average proportion of high-cost individuals, such as young and healthy enrollees (i.e. a low-risk portfolio). The total amount of money in the RE Fund is allocated among all insurers.

Figure A.1: RE scheme in the Netherlands



§ A.1.2 Which risk adjusters are included?

A first important choice about the design of the RE-model is which risk adjusters are included. The Dutch RE-model of 2014, which is the most recent RE-model studied in this thesis, uses an advanced set of risk adjusters, including four morbidity-based risk adjusters. Table A.1 describes the risk adjusters included in this model. For each risk adjuster, dummy variables are defined for the risk classes, resulting into in total 132 dummy variables for the Dutch RE-model of 2014.

The definition of the risk adjusters somewhat changed over the study period 2006 to 2011: e.g. an extra pharmacy-based cost group (PCG) and/or diagnostic cost group (DCG). These small changes in the definition of the risk adjusters did not lead to significant changes in the predictive performance of the RE-model. Two important changes during the study period are the inclusion of the multiple year high cost groups (MHC-groups) and durable medical equipment groups (DME-groups), which led to a significant increase in model's predictive performance, especially the MHC-groups.

§ A.1.2 Which healthcare expenses are compensated?

A second important choice on the design of the RE-model is which expenses are compensated. In the Netherlands, just as is the case in all countries with RE, *total observed expenses* are those expenses to be compensated and so, total observed expenses is the dependent variable. In model estimation, total observed expenses are annualized and a weight is used for the enrolment period.

Total observed expenses are those expenses related to the services included in the basic benefit package in the respective year. Throughout this thesis, all expenses included in the basic benefit package per year were used, except (long-term) mental healthcare expenses. This is because for these expenses another RE-model with some other risk adjusters are used in the Netherlands. Insurers bear financial risk for a part or the total amount of expenses related to the services included in the basic benefit package. For those expenses for which insurers do not bear 100% financial risk, they are compensated ex-post the contract period

Table A.1: Definition of the risk adjusters in the Dutch RE-model of 2014

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 years, 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. Information on age and gender are derived from insurers' administrative files.
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 non-western immigrants, percentage of one-person households, distance to the general practitioner, distance to a hospital, and degree of urbanization. This information is available at the regional level; and not at the individual level. The clustering of zip codes is determined as follows: (1) estimating the RE-model without region; (2) calculating residual expenses of the model in step one; (3) regressing residual expenses as calculated in step 2 on the information at the zip code level; (4) estimate residual expenses of the model in the third step; (5) determine the thresholds of residual expenses for obtaining ten clusters, each containing approximately 10% of the total observations.
Source of income * Age (18 risk classes): 4 categories of source of income (disability benefits, social security benefits, self-employed, and other), 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. Information on source of income is derived from an agency collecting tax revenues and the registration agency for social benefits and assistance.
Pharmacy-based cost groups (PCGs, 24 risk classes): Individuals are assigned to a PCG when they have used more than 180 defined daily dosages of specific prescribed drugs 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 the relevant drugs in the previous year are 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 in only one of these PCGs. The pharmacy information is derived from a national database on drug prescriptions.
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 without a hospital admission are classified in DCG 0. Individuals can be classified into only one DCG, the one with the highest follow-up costs. The DCGs are based on hospital diagnoses, whereby different diagnoses are classified into homogeneous cost groups. The information is derived from a national database on hospital declarations.
Socioeconomic status * Age (12 risk classes): 4 socio-economic classes: SES 0 is for individuals living on a home address with more than 15 persons (i.e. residential care 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. The information on income and households is derived from an agency collecting tax revenues.
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 are not classified in one of these risk classes, i.e. those who do not have high expenses in three prior years, are classified in a separate risk class (MHC 0). The information on prior years' expenses are derived from insurers' administrative files.
Durable medical equipment groups (DME, 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 use insulin pumps, DME 2 is for those who use a catheter, DME 3 is for those who use a colostomy, and DME 4 is for those who use a trachea colostomy. Individuals who have not used durable medical equipment are classified in a separate risk class (DME 0). Information on the usage of durable medical equipment is derived from insurers' administrative files.

via risk sharing arrangements. This thesis exclusively focusses on the *ex-ante* RE-model and not on risk sharing arrangements that were enforced during the study period in the Netherlands.

Time lag between data collection and estimation of the risk equalization model

Since it takes time to collect all relevant information for each individual in the population, the most recent dataset that can be used to estimate the actual RE-model in a given year is data from at least three years prior to the estimation-year of the RE-model. This delay is mainly caused by the time that it takes to obtain certain hospital declarations. For example, the Dutch RE-model of 2014 is estimated on the administrative dataset from 2011, which includes cost and demographic information from 2011, diagnostic information in terms of the PCGs, DCGs, and DME-groups from 2010, while the MHC-groups are based on cost information from 2010, 2009, and 2008.

§ A.1.3 How to specify the RE-model?

A third important choice on the design of the RE-model is the specification of the model. The Dutch RE-model is estimated by OLS on untransformed data, which is the conventional estimation technique for RE-models. Two important reasons why OLS on untransformed data is used for estimating RE-models are: (i), OLS is easier to use and interpret than alternative estimation techniques, such as two-part models, generalized linear models, and/or models based on (log-) transformed data. In the context of RE, transparency and commitment under all relevant stake-holders; e.g. the regulator, policymakers, and insurance companies, are highly important; and (ii), some studies have shown that OLS may provide a similar model fit as more advanced estimation techniques when the sample size is large (enough), which may be millions of observations (Dunn, 2003; Jones, 2010; Mihaylova et al., 2011; Powers et al., 2005).

GENERAL APPENDIX 2: DEFINITION OF THE EVALUATION-GROUPS

Table A.2: Definition of the pre-defined selective groups used for evaluating models' predictive performance

Group	Definition
General health status is poor	The following question is answered with "bad or "very bad": "How do you rate your health status?"
At least one long-term disease	The following question is answered with "yes": "Do you have one or more long-term diseases?"
Obesity	Obesities according to the Quetelet index, individuals with a BMI > 30.
OECD limitations in hearing	At least one of the following questions is answered with "yes, but with many difficulties" or "no, I cannot": "Can you follow a conversation in a group of three or more persons?"; "Can you hold a conversation with another person?"
OECD limitations 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 limitations in moving	At least one of the following three questions is answered with "yes, but with many difficulties" or "no, I cannot": "Can you lift a weight of 5 kilo's for 10 meters?"; "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 limitations in talking	The following question is answered with "yes, but with many difficulties" or "no, I cannot": "Can you speak intelligible?"
OECD limitations in eating	The following question is answered with "yes, but with many difficulties" or "no, I cannot": "Can you bite and chew hard food?"
The lowest score on physical health scales	Individuals with the worst (lowest 10%) on the SF-12 physical component summary scale[a.]
A low score on physical health scales	Individuals with a bad score (lowest 20%) on the SF-12 mental component summary scale[a.]
The lowest score on mental health scales	Individuals with the worst (lowest 10%) on the SF-12 mental component summary scale[a.]
A low score on mental health scales	Individuals with a bad score (lowest 20%) on the SF-12 mental component summary scale[a.]
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 A.2: (continued)

Group	Definition
Self-reported diseases (19 groups)	A "yes" on the question: 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 regularly periods of dizziness 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?
Two self-reported diseases / Three or more self-reported diseases	Two times "Yes" or Three or more times "Yes" on the questions about the self-reported disease or disorder.
Contact general practitioner in the past year	A "Yes" on the question: "Did you contact the general practitioner in the last 12 months?"
Contact medical specialist in the past year	A "Yes" on the question: "Did you contact the medical specialist in the last 12 months?"
Hospitalization in the past year	A "Yes" on the question: "Did you had a hospital admission in the last 12 months?"
Contact physiotherapist in the past year	A "Yes" on the question: "Did you contact the physiotherapist the last 12 months?"
Prescribed drugs use in the past 14 days	A "Yes" on the question: "Did you use prescribed drugs during the last 14 days?"
Glasses or contact lenses	A "Yes" on the question: "Do you use glasses or contact lenses?"
Hearing-aid	A "Yes" on the question: "Do you use a hearing-aid?"
Complete dentures	A "Yes" on the question: "Do you use complete dentures?"
Use of durable medical equipment	At least one of the following questions is answered with "always": "How many times do you use an aid for walking (walker)?"; "How many times do you use a wheelchair (hand or electronic)?"; "How many times do you use an orthopedic shoe?"; "How many times do you use a prosthesis (arm or leg)?"; "How many times 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?"
Contact with home nurse in the past year[b]	A "Yes" on the question: "Did you contact a home nurse in the last 12 months?"
Contact with home care practitioner in the past year	A "Yes" on the question: "Did you contact a home care practitioner in the last 12 months?"
Home help assistance in the past year[b]	A "Yes" on the question: "Do you have professional assistance with activities at home in the last 12 months?"

Footnotes Table A.2:

a. Ware, J.E. Jr., Kosinski, M., Keller, S.D. (1996). A 12-Item Short-Form Health Survey: Construction of Scales and Preliminary tests of Reliability and Validity. *Medical Care*. 34, 220-233.

b. The same health survey is conducted over time and therefore, the same definition of the evaluation-groups could be used. Since 2010, it was possible to define one additional group: i.e. the 'home help assistance'-group was divided into a 'home help'-group and 'home nursing'-group. Consequently, on 2008-data 45 groups were defined and on 2010-data 46 groups.

Abbreviations

LIST OF ABBREVIATIONS

Abbreviation	Definition	
CPM	Cumming's Prediction Measure	A measure-of-fit for risk equalization models.
C-type risk factor	Compensation-type risk factor	Risk factor for which the regulator desires compensation: e.g., individuals' health status; also called S-type risk factor.
DCG	Diagnostic Cost Group	A risk adjuster included in the Dutch risk equalization model since 2004.
DME-group	Durable Medical Equipment-Group	A risk adjuster included in the Dutch risk equalization model since 2014.
GLM	Generalized Linear Model	A statistical model specification.
MAPE	Mean Absolute Prediction Error	A measure-of-fit for risk equalization models.
MHC-group	Multiple-Year High Cost Group	A risk adjuster included in the Dutch risk equalization model since 2012.
MPE	Mean Prediction Error	A measure-of-fit for risk equalization models.
MSPE	Mean Squared Prediction Error	A measure-of-fit for risk equalization models.
N-type risk factor	Non-subsidy-type risk factor	Risk factor for which the regulator does not desire compensation: e.g., inefficiency in provision of services; also called R-type risk factor.
OLS	Ordinary Least Squares	A statistical model specification, which is the conventional estimation technique in the field of risk equalization.
PCG	Pharmacy-based Cost Group	A risk adjuster included in the Dutch risk equalization model since 2000.
PR	Predictive Ratio	A measure-of-fit for risk equalization models.
R^2	R-squared	A measure-of-fit for risk equalization models.
RE Fund	Risk Equalization Fund	The fund is filled with mandatory contributions from taxes, insurers, or consumers. The money in this fund is allocated among insurers by means of risk-adjusted payments.
RE-model	Risk Equalization model	The prediction model that is used to calculate the risk-adjusted payments.
R-type risk factor	Responsibility-type risk factor	Risk factor for which the regulator does not desire compensation: e.g., inefficiency in provision of services; also called N-type risk factor.
SES	Socioeconomic-status	Information about the income per household that is used to define a risk adjuster in the Dutch risk equalization model.
SF-12	12-item Short Form survey	A generic short-form survey with 12 questions selected from the SF-36 health survey. The questionnaire includes scales of mental and physical functioning and overall health-related quality of life.
S-type risk factor	Subsidy-type risk factor	Risk factor for which the regulator desires compensation: e.g., individuals' health status; also called C-type risk factor.
2PMs	Two-part models	A statistical model specification.

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.S., 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.
* Ash, A., Byrne-Logan, S. (1998). How well do models work? Predicting health care costs. Proceedings of the Section on Statistics in Epidemiology of the American Statistical Association: 42-49.
* Ash, A.S., Ellis R.P., Pope, G.C., Ayanian, J.Z., Bates, D.W., Burstin, H., Iezzoni, L.I., MacKay, E., Yu, W. (2000). Using Diagnosis to describe populations and predict expenses. *Health Care Financing Review, 21*(3), 7-28.
* Ash, A.S., McCall, N., Fonda, J., Hanchate, A., Speckman, J. (2005). Risk assessment of Military Populations to predict health care expenses and utilization. Final report: Research Triangle Institute, Washington, DC and Boston University School of Medicine, Boston.
* Ash, A.S., Ellis, R.P. (2012). Risk-adjusted payment and performance assessment for primary care. *Medical Care, 50*(8), 643-653.
* Babyak, M.A. (2004). What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. *Psychosomatic Medicine, 66*, 411-421.
* Barry, C.L., Weiner, J.P., Lemke, K., Busch, S.H. (2012). Risk adjustment in health insurance exchanges for individuals with mental illness. *American Journal of Psychiatry, 169*(7), 704-709.
* Baumgartner, C., Busato, A. (2012). Risikoselektion in der Grundversicherung. *Schweizerische Ärztezeitung. 93*(13), 510-513.
* Basu, A., Manning, W.G. (2009). Issues for the Next Generation of Health Care Costs Analyses. *Medical Care, 47*, S109-S114.
* Beck, K, Zweifel, P. (1998). Cream-skimming in deregulated social health insurance: evidence from Switzerland. In: Zweifel, P. (Ed.), Health, the Medical Profession, and Regulation. Kluwer, Dordrecht, pp. 211-227.
* Beck, K. (2000). Growing importance of capitation in Switzerland. *Health Care Management Science, 3*(2), 111-119.
* Beck, K., Spucher, S., Holly, A., Gardiol, L. (2003). Risk Adjustment in Switzerland. *Health Policy. 65*, 63-74.
* Beck, K., Trottmann, M., Zweifel, P. (2010). Risk adjustment in health insurance and its long-term effectiveness. *Journal of Health Economics, 29*, 489-498.
* 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.
* Breyer, F., Heineck, M., Lorenz, N. (2003). Determinants of health care utilization by German sickness fund members - with application to risk adjustment. *Health Economics, 12*(5), 367-376.
* Breyer, F., Bundorf, M.K., Pauly, M.V. (2012). *Health Care Spending, and Payment to Health Plans*. In: Pauly, M.V., McGuire, T.G., and Barros, P.P. Handbook of Health Economics, pp. 691-762. Elsevier Science B.V., Amsterdam.
* Buchner, F., Goepffarth, D., Wasem, J. (2013). The new risk adjustment formula in Germany: implementation and first experiences. *Health Policy*, *109*, 253-262.

* Buchner, F., Wasem, J., Schillo, S. (2014). Regression trees in the German risk adjustment formula. *Working paper,* version 23 September 2014.
* Buntin, M.B., Garber, A.M., McClellan, M., Newhouse, J.P. (2004). The expenses of decedents in the Medicare program: implications for payments to Medicare+Choice plans. *Health Services Research, 39*(1), 111-130.
* Buntin, M.B., Zaslavsky, A.M. (2004). Too much abo about two-part models and transformation? Comparing methods of modelling Medicare expenditures. *Journal of Health Economics, 23*, 525-542.
* Calderón-Larrañaga, A., Abrams, C., Poblador-Plou, B., Weiner, J.P., Prados-Torres, A. (2010). Applying diagnosis and pharmacy-based risk models to predict pharmacy use in Aragon, Spain: the impact of a local calibration. *BMC Health Services Research, 10*(22).
* Cameron, A.C., and Windmeijer, F.A.G. (1996). R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics, 14*(2), 209-220.
* Carter, G.M., Bell, R.M., Dubois, R.W., Goldberg, G.A, Keeler, E.B., McAlearney, J.S., Post, E.P., Rumpel, J.D. (2000). A clinically detailed risk information system for cost. *Health Care Financing Review, 21*(3), 65-91.
* Chalupka, R. (2010). Improving risk adjustment in the Czech Republic. Prague Economic Papers, 3, 236-250.
* Chang, H., Weiner, J.P. (2010). An in-depth assessment of a diagnosis-based risk adjustment model based on national health insurance claims: the application of the Johns Hopkins Adjusted Clinical Group case-mix system in Taiwan. *BMC Medicine, 8*(7), 1-13.
* Chang, R., Lin, W., Hsieh, C., Chiang, T. (2002). Healthcare utilization patterns and risk adjustment under Taiwan's national health insurance system. *Journal of the Formosan Medical Association, 101*(1), 52-59.
* Cumming, R.B., Knutson, D., Cameron, B.A., Derrick, B. (2002). A comparative analysis of claims-based methods of health risk assessment for commercial populations. Society of Actuaries, USA.
* DeSalvo, K.B., Jones, T.M., Peabody, J., McDonald, J., Fihn, S., Fan, V., He, J., Muntner, P. (2009). Health care expenses prediction with a single item, self-rated health measure. *Medical Care, 47*(4), 440-447.
* Donato, R., Richardson, J. (2006). Diagnosis-based risk adjustment and Australian health system policy. *Australian Health Review, 30*(1), 83-99.
* Duan, N., Manning, W.G., Morris, C.N., Newhouse, J.P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics, 1*(2), 115-126.
* Duckett, S.J., and Agius, P.A. (2002). Performance of diagnosis-based risk adjustment measures in a population of sick Australians. *Australian and New Zealand Journal of Public Health, 26*(6), 500-507.
* Dudley, R.A., Medlin, C.A., Hammann, L.B., Cisternas, M.G., Brand, R., Rennie, D.J., Luft, H.S. (2003). The best of both worlds? Potential of hybrid prospective/concurrent risk adjustment. Medical Care, 41(1), 56-69.
* Dunn, G., Mirandola, M., Amaddeo, F., Tansella, M. (2003). Describing, explaining or predicting mental health care costs: a guide to regression models: Methodological review. *The British Journal of Psychiatry, 183*, 398-404.
* Eggleston, K., Bir, A. (2009). Measuring Selection Incentives in Managed Care: Evidence from the Massachusetts State Employee Insurance Program. *Journal of Risk and Insurance, 76*(1), 159-175.
* Eijkenaar, F., van Kleef, R.C., van Veen, S.H.C.M., van Vliet, R.C.J.A. (2013). Onderzoek risicovereveningsmodel 2014: berekening normbedragen. WOR 658. Onderzoek voor het Ministerie van Volksgezondheid, Welzijn en Sport. Instituut Beleid & Management Gezondheidszorg, Erasmus Universiteit Rotterdam.

* Eijkenaar, F., van Kleef, R.C., van Veen, S.H.C.M., van Vliet, R.C.J.A. (2015). Onderzoek risicovereveningsmodel 2016: berekening normbedragen. WOR 749. Onderzoek voor het Ministerie van Volksgezondheid, Welzijn en Sport. Instituut Beleid & Management Gezondheidszorg, Erasmus Universiteit Rotterdam.
* Ellis, R.P., Ash, A.S. (1995). Refinements to the Diagnostic Cost Group (DCG) Model. *Inquiry, 32*, 418-429.
* Ellis, R.P., Mookim, P.G. (2009). Cross-validation methods for risk adjustment models. Working paper.
* Enthoven, A. (1988). Managed competition of alternative delivery systems. *Journal of Health Politics, Policy and Law, 13*(2): 305.
* Ettner, S.L., Frank, R.G., Mark, T., Smith, M.W. (2000). Risk adjustment of capitation payments to behavioral health care carve-outs: how well do existing methodologies account for psychiatric disability? *Health Care Management Science, 3*(2), 159-169.
* Ettner, S.L., Frank, R.G., McGuire, T.G., Hermann, R.C. (2001). Risk adjustment alternatives in paying for behavioral health care under Medicaid. *Health Services Research, 36*(4), 793-811.
* 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.
* Fleishman, J.A., Cohen, J.W., Manning, W.G., Kosinski, M. (2006). Using the SF-12 health status measure to improve predictions of medical expenses. *Medical Care, 44*(5 suppl.), I-54- I-63.
* Fox, J. (2008). *Applied regression analysis and generalized linear models*. Second edition, Sage Publications, Inc.
* Frank, R.G., Glazer, J., McGuire, T.G. (1998). Measuring adverse selection in managed health care. NBER working paper series, working paper 6825, national bureau of economic research, Cambridge M.A.
* Frogner, B.K., Anderson, G.F., Cohen, R.A., Abrams, C. (2011). Incorporating new research into Medicare risk adjustment. *Medical Care, 49*(3), 295-300.
* Gail, M., Krickeberg, K., Samet, J., Tsiatis, A., Wong, W. (2009). *Statistics for Biology and Health*. New York: Springer Science + Business Media, LLC.
* Garber, A.M., Macurdy, T.E., McClellan, M.B. (1998). Persistence of Medicare Expenditures Among Elderly Beneficiaries. In: Garber, A.M. *Frontiers in Health Policy*, pp. 153-180.
* García-Goñi, M., Ibern, P. (2008). Predictability of drug expenses: an application using morbidity data. *Health Economics, 17*(1), 119-126.
* García-Goñi, M., Ibern, P., Inoriza, J.M. (2009). Hybrid risk adjustment for pharmaceutical benefits. *European Journal of Health Economics, 10*(3), 299-308.
* Geruso, M., McGuire, T.G. (2014). Tradeoffs in the design of health plan payments systems: fit, power and balance. *NBER working paper series*, working paper 20359, National Bureau of Economic Research, Cambridge., MA.
* Gilmer, T., Kronick, R., Fishman, P., Ganiats, T.G. (2001). The Medicaid Rx model: pharmacy-based risk adjustment for public programs. *Medical Care, 39*(11), 1188-1202.
* Hanley, G.E., Morgan, S., Reid, R.J. (2010). Explaining prescription drug use and expenses using the Adjusted Clinical Groups case-mix system in the population of British Columbia, Canada. *Medical Care, 48*(5), 402-408.
* Hastie, T.R., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer. Second Edition.
* Hornbook, M.N., Goodman, M.J. (1996). Chronic Disease, Functional Health Status, and Demographics: A Multi-Dimensional Approach to Risk Adjustment. *Health Services Research, 31*(3), 283-307.

* Hsu, J., Huang, J., Fung, V. Price, M., Brand, R., Hui, R., Fireman, B., Dow, W.H., Bertko, J., Newhouse, J.P. (2009). Distributing $800 billion: an early assessment of Medicare part D risk adjustment. *Health Affairs, 28*(1), 215-225.
* Hsu, J., Fung, V., Huang, J., Price, M., Brand, R., Hui, R., Fireman, B., Dow, W.H., Bertko, J., Newhouse, J.P. (2010). Fixing flaws in Medicare drug coverage that prompts insurers to avoid low-income patients. *Health Affairs, 29*(12), 2335-2342.
* 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.
* Hwang, W., Ireys, H.T., Anderson, G.F. (2001). Comparison of risk adjusters for Medicaid-enrolled children with and without chronic health conditions. *Ambulatory Pediatrics, 1*(4), 217-224.
* Jones, A.M. (2010). Models for health care. Working paper, University of York. http://www.york.ac.uk/res/herc/documents/wp/10_01.pdf. Accessed 13 June 2013.
* Kanters, T.A., Brouwer, W.B.F., van Vliet, R.C.J.A., van Baal, P.H.M., Polder, J.J. (2013). A new prevention paradox: the trade-off between reducing incentives for risk selection and increasing the incentives for prevention for health insurers. *Social Sciences & Medicine*, *76*, 150-158.
* Kapur, K., Young, A.S., Murata, D. (2000). Risk adjustment for high utilizers of public mental health care. *The Journal of Mental Health Policy and Economics, 3*(3), 129-137.
* Kautter, J., Ingber, M., Pope, G.C. (2008). Medicare risk adjustment for the frail elderly. *Health Care Financing Review, 30*(2), 83-93.
* Kautter, J., Ingber, M., Pope, G.C., Freeman, S. (2012). Improvements in Medicare part D risk adjustment: beneficiary access and payment accuracy. *Medical Care, 50*(12), 1102-1108.
* 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.
* 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.
* Kuhlthau, K., Ferris, T.G., Davis, R.B., Perrin, J.M., Iezzoni, L.I. (2005). Pharmacy- and diagnosis-based risk adjustment for children with Medicaid. *Medical Care, 43*(11), 1155-1159.
* Lamers, L.M., van Vliet, R.C.J.A. (1996). Multiyear diagnostic information from prior hospitalizations as a risk-adjuster for capitation payments. *Medical Care, 34*, 549-561.
* Lamers, L.M. (1997). Capitation payments to competing Dutch sickness funds based on diagnostic information from prior hospitalizations. Ph.D. Dissertation, Erasmus University Rotterdam, Rotterdam.
* Lamers, L.M. (2001). Health-based risk adjustment: Is inpatient and outpatient diagnostic information sufficient? *Inquiry, 38*(4), 423-431.
* 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.
* Lamers, L.M., van Vliet, R.C.J.A. (2004). The pharmacy-based cost group model: validating and adjusting the classification of medications for chronic conditions to the Dutch situation. *Health Policy, 68*(1), 113-121.
* Last, M., Maimon, O., Minkov, E. (2002). Improving stability of decision trees. *International Journal of Pattern Recognition and Artificial Intelligence*, *16*, 2.
* Levy, J.M., Robst, J., Ingber, M.J. (2006). Risk-adjustment system for the Medicare capitated ESRD program. *Health Care Financing Review, 27*(4), 53-69.

* Loh, H.-W., Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*, 815-840.
* Loh, H.-W. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, *12*, 361-386.
* Luft, H.S., Dudley, R.A. (2004). Assessing risk-adjustment approaches under non-random selection. *Inquiry, 41*(2), 203-217.
* Maciejewski, M.L., Liu, C.–F., Derleth, A., McDonell, M., Anderson, S., Fihn, S.D. (2005). The performance of administrative and self-reported measures for risk adjustment of veterans affairs expenses. *Health Services Research, 40*(3), 887-904.
* Maciejewski, M.L., Liu, C.–F., Fihn, S.D. (2009). Performance of comorbidity, risk adjustment, and functional status measures in expenses prediction for patients with diabetes. *Diabetes Care, 32*(1), 75-80.
* Madden, C.W., Mackay, B.P., Skillman, S.M., Ciol, M., Diehr, P.K. (2000). Risk adjusting capitation: applications in employed and disabled populations. *Health Care Management Science, 3*(2), 101-109.
* Manning, W.G., Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Journal of Health Economics, 20*, 461-494.
* Manning, W.G., Basu, A., Mullahy, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics, 24*, 465-488.
* Mark, T.L., Ozminkowski, R.J., Kirk, A., Ettner, S.L., Drabek, J. (2003). Risk adjustment for people with chronic conditions in private sector health plans. *Medical Decision Making, 23*(5), 397-405.
* McIntyre, S.H., Montgomery, D.B., Srinivasan, V., Weitz, B.A. (1983). Evaluating the Statistical Significance of Models Developed by Stepwise Regression. *Journal of Marketing Research, 20*, 1-11.
* Mihaylova, B., Briggs, A., O'Hagan, A., Thompson, S.G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics, 20*, 879-916.
* Monheit, A.C. (2003). Persistence in Health Expenditures in the Short Run: Prevalence and Consequences. *Medical Care, 41*, 7, 53-64.
* Newhouse, J.P., Manning, W.G., Keeler, E.B., Sloss, E.M. (1989). Adjusting capitation rates using objective health measures and prior utilization. *Health Care Financing Review, 10*(3), 41-54.
* Newhouse, J.P. (1996). Reimbursing health plans and health providers: efficiency in production versus selection. Journal of Economic Literature, volume XXXIV, 1236-1263.
* Newhouse, J.P., Price, M.P., Hung, J., McWilliams, J.M., and Hsu, J. (2012). Steps to reduce favorable risk selection in Medicare advantage largely succeeded, boding well for health insurance exchanges. *Health Affairs, 13*(12), 2618-2628.
* Newhouse, J.P., Price, M., McWilliams, J.M., Hsu, J., McGuire, T.G. (2015). How much favorable selection is left in Medicare Advantage? *American Journal of Health Economics, 1*(1), 1-26.
* Noyes, K., Liu, H., Temkin-Greener, H. (2006). Cost of caring for Medicare beneficiaries with Parkinson's disease: Impact of the CMS-HCC risk-adjustment model. *Disease Management, 9*(6), 339-348.
* Oates, T., Jensen, D. (1997). The effects of training set size on decision tree complexity. *Machine Learning: Proceedings of the Fourteenth International Conference*. Morgan Kaufmann, 254-262.
* Pacala, J.T., Boult, C., Urdangarin, C., McCaffrey, D. (2003). Using self-reported data to predict expenses for the health care of older people. *Journal of the American Geriatrics Society, 51*(5), 609-614.
* 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.
* Pietz, K., Ashton, C.M., McDonell, M., Wray, N.P. (2004). Predicting healthcare expenses in a population of veterans affairs beneficiaries using diagnosis-based risk adjustment and self-reported health status. *Medical Care, 42*(10), 1027-1035.

* Pindyck, R.S., Rubinfeld, D.L. (1998). *Econometric models and economic forecasts*. The McGraw-Hill Companies, Inc.
* 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. (2000a). Principal inpatient diagnostic cost group model for Medicare risk adjustment. *Health Care Financing Review, 21*(3), 93-118.
* Pope, G.C., Ellis, R.P., Ash, A.S., Ayanian, J.Z., Bates, D.W., Burstin, H., Iezzoni, L.I., Marcantonio, E., Wu, B. (2000b). 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., Kautter, J., Ellis, R.P., Ash, A.S., Ayanian, J.Z., Iezzoni, L.I., Ingber, M.J., Levy, J.M., Robs, J. (2004). Risk adjustment of Medicare capitation payments using the CMS-HCC model. *Health Care Financing Review, 25*(4), 119-141.
* Powers, C.A., Meyer, C.M., Roebuck, M.C., Vaziri, B. (2005). Predictive modeling of total healthcare expenses using pharmacy claims data: A comparison of alternative econometric cost modeling techniques. *Medical Care, 43*(11), 1065-1072.
* Prinsze, F.J., van Vliet, R.C.J.A. (2007). Health-based risk adjustment: Improving the pharmacy-based cost group model by adding diagnostic cost groups. *Inquiry, 44*(4), 469-480.
* Reid, R.J., MacWilliam, L., Verhulst, L., Roos, N., Atkinson, M. (2001). Performance of the ACG case-mix system in two Canadian provinces. *Medical Care, 39*(1), 86-96.
* Rein, D.B. (2005). A matter of classes: stratifying health care populations to produce better estimates of inpatient expenses. *Health Services Research, 40*(4), 1217-1233.
* Riley, G.F. (2000). Risk adjustment for health plans disproportionately enrolling frail Medicare beneficiaries. *Health Care Financing Review, 21*(3), 135-148.
* Robinson, J.W., Karon, S.L. (2000). Modeling Medicare expenses of PACE populations. *Health Care Financing Review, 21*(3), 149-170.
* 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.
* Robst, J. (2009). Development of a Medicaid behavioral health case-mix model. *Evaluation Review, 33*(6), 519-538.
* Robst, J., Levy, J.M., Ingber, M.J (2007). Diagnosis-based risk adjustment for Medicare prescription drug plan payments. *Health Care Financing Review, 28*(4), 15-30.
* Sales, A.E., Liu, C.F, Sloan, K.L., Malkin, J., Fishman, P.A., Rosen, A.K., Loveland, S., Paul Nichol, W., Suzuki, N.T., Perrin, E., Sharp, N.D., Todd-Stenberg, J. (2003). Predicting expenses of care using a pharmacy-based measure risk adjustment in a veteran population. *Medical Care, 41*(6), 753-760.
* Sarma, K.S. (2007). *Predictive Modeling with SAS® Enterprise Miner ™: Practical Solutions for Business Applications*. Cary, NC: SAS® Institute Inc.
* Schokkaert, E., van de Voorde, C. (2004). Risk selection and the specification of the conventional risk adjustment formula. *European Journal of Health Economics, 23*, 1237-1259.
* Schokkaert, E., van de Voorde, C. (2006). Incentives for risk Selection and omitted variables in the risk adjustment formula. *Annals of Economics and Statistics, 83/84*, 327-351.
* Schokkaert, E., van de Voorde, C. (2009). Direct versus indirect standardization in risk adjustment. *Journal of Health Economics, 28*, 361-374.
* Schut, F.T., van de Ven, W.P.M.M. (2010). Uitvoering AWBZ door zorgverzekeraars onverstandig. *Economisch Statistische Berichten, 95*(4591), 486-489.
* Shen, Y., Ellis, R.P. (2002). How profitable is risk selection? A comparison of four risk adjustment models. *Health Economics, 11*(2), 165-174.

* Shih, Y.-S., Tsia, H.-W. (2004). Variable selection bias in regression trees with constant fits. *Computational Statistics and Data Analysis*, *45*, 595-607.
* Shmueli, A., Messika, D., Zmora, I., Oberman, B. (2010). Health care expenses during the last 12 months of life in Israel: estimation and implications for risk-adjustment. *International Journal of Health Care Finance and Economics, 10*(3), 257-273.
* Shmueli, A., Nissan-Engelcin, E. (2013). Local availability of physicians' services as a tool for implicit risk selection. *Social Science & Medicine, 84*: 53-60.
* Shmueli, A. (2015). On the calculation of the Israeli risk adjustment rates. *European Journal of Health Economics, 16*, 271-277.
* Stam, P.J.A., van de Ven, W.P.M.M. (2006). Risicoverevening in de zorgverzekering: Een evaluatie en oplossingsrichtingen voor verbetering. Research Report, iBMG, Erasmus University Rotterdam, Rotterdam.
* Stam, P.J.A. (2007). Testing the effectiveness of risk equalization models in health insurance. Ph.D. Dissertation, Erasmus University Rotterdam, Rotterdam.
* Stam, P.J.A., van de Ven, W.P.M.M. (2008). De harde kern in risicoverevening. *Economisch Statistische Berichten, Februari*, 104-107.
* Stam, P.J.A., van Vliet, R.C.J.A., van de Ven, W.P.M.M. (2010a). A limited-sample benchmark approach to assess and improve the performance of risk equalization models. *Journal of Health Economics, 29*, 426-437.
* Stam, P.J.A., van Vliet, R.C.J.A., van de Ven, W.P.M.M. (2010b). 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.
* Temkin-Greener, H., Meiners, M. R., Gruenberg, L. (2001). PACE and the Medicare+Choice risk-adjusted payment model. *Inquiry, 38*(1), 60-72.
* Thompson, M.L. (1978). Selection of variables in Multiple Regression: Part 1. A Review and Evaluation. *International Statistical Review, 46*, 1-19.
* van Barneveld, E.M., Lamers, L.M., van Vliet, R.C.J.A., van de Ven, W.P.M.M. (2000). Ignoring small predictable profits and losses: A new approach for measuring incentives for cream skimming. *Health Care Management Science, 3*, 131-140.
* van Barneveld, E.M., Lamers, L.M., van Vliet, R.C.J.A., van de Ven, W.P.M.M. (2001). Risk sharing as a supplement to imperfect capitation: a trade-off between selection and efficiency. *Journal of Health Economics, 20*, 147-168.
* van Kleef, R.C., van Vliet, R.C.J.A. (2010). Prior use of durable medical equipment as a risk adjuster for health-based capitation. *Inquiry, 47*(4), 343-358.
* van Kleef, R.C., van Vliet, R.C.J.A. (2012). Improving risk equalization using multiple-year high cost as a health indicator. *Medical Care, 50*(2), 140-144.
* van Kleef, R.C., van Vliet, R.C.J.A., van de Ven, W.P.M.M. (2012a). Risicoverevening tussen zorgverzekeraars: Kwantificering modelverbeteringen 1993-2011. *Tijdschrift voor Gezondheidswetenschappen. 90*(5), 312-326
* van Kleef, R.C., van Vliet, R.C.J.A., van de Ven, W.P.M.M. (2012b). Risicoverevening 2012. Een analyse van voorspelbare winsten en verliezen op subgroepniveau. Onderzoeksrapport, Erasmus Universiteit Rotterdam, iBMG, Rotterdam.

* van Kleef, R.C., van Vliet, R.C.J.A., van de Ven, W.P.M.M. (2012c). Diagnosis-based cost groups in risk adjustment: The effects of including outpatient diagnoses. Research report, iBMG, Erasmus University Rotterdam, Rotterdam.
* van Kleef, R.C., van de Ven, W.P.M.M., van Vliet, R.C.J.A. (2013a). Risk selection in a regulated health insurance market: a review of the concept, possibilities, and effects. *Expert Review of Pharmaeconomics & Outcomes Research, 13*(6), 743-752.
* van Kleef, R.C., van Vliet, R.C.J.A., van de Ven, W.P.M.M. (2013b) Risk Equalization in the Netherlands: an empirical evaluation. *Review of Pharmaeconomics & Outcomes Research, 13*(6), 829-839.
* van Kleef, R.C., van Vliet, R.C.J.A., van de Ven, W.P.M.M. (2014). Risicoverevening 2014 voor de somatische zorg. Analyse van uitkomsten op subgroepniveau. *Research report*, Erasmus Universiteit Rotterdam, iBMG, Rotterdam.
* van Kleef, R.C., McGuire, T.G., van Vliet, R.C.J.A., van de Ven, W.P.M.M. (2015). Improving risk equalization with constrained regression. NBER Working paper series, national bureau of economic research, NBER Working paper No. 21570.
* van Veen, S.H.C.M., van Kleef, R.C, van de Ven, W.P.M.M., van Vliet, R.C.J.A. (2015a). 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.
* van Veen, S.H.C.M., van Kleef, R.C, van de Ven, W.P.M.M., van Vliet, R.C.J.A. (2015b). Improving the prediction model used in risk equalization: cost and diagnostic information from multiple prior years. *European Journal of Health Economics, 16*(2), 201-218.
* van de Ven, W.P.M.M., Ellis, R.P. (2000). *Risk adjustment in competitive health plan markets*. In Handbook of health economics. Cutler, A. and J.P. Newhouse. First edition, pp. 755-845. Elsevier Science, B.V., Amsterdam.
* van de Ven, W.P.M.M., 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.
* van de Ven, W.P.M.M., Schut, F.T. (2011). 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.
* van de Ven, W.P.M.M., Beck, K., Buchner, F., Schokkaert, E., Schut, F.T., Shmueli, A., Wasem, J. (2013). Preconditions for efficiency and affordability in competitive healthcare markets: are they fulfilled in Belgium, Germany, Israel, the Netherlands, and Switzerland? *Health Policy, 109*(3), 226-245.
* van Vliet, R.C.J.A., van de Ven, W.P.M.M. (1992). Towards a capitation formula for competing health insurers: an empirical analysis. *Social Sciences and Medicine, 34*, 1035-1048.
* van Vliet, R.C.J.A., van de Ven, W.P.M.M. (1993). Capitation payments based on prior hospitalizations. *Health Economics, 2*, 177-188.
* van Vliet, R.C.J.A. (2006). Free choice of health plan combined with risk-adjusted capitation payments: are switchers and new enrollees good risks? *Health Economics, 15*(8), 763-774.
* van Vliet, R.C.J.A., Everhardt, T.P., van Asselt, M.M., Goudriaan, R., Mazzola, G.J. (2011). Overall Toets risicovereveningsmodel somatische zorg 2012, WOR 578, Eindrapportage, Den Haag: APE.
* van Vliet, R.C.J.A., van Asselt, M.M., Goudriaan, R., Mazzola, G.J., Everhardt, T.P. (2012). Overall Toets risicovereveningsmodel somatische zorg 2013, WOR 625, Eindrapportage, Den Haag: APE.
* Vargas, V., Wasem, J. (2006). Risk adjustment and primary health care in Chile. *Croatian Medical Journal, 47*(3), 459-468.
* Veazie, P.J., Manning, W.G., Kane, R.L. (2003). Improving risk adjustment for Medicare capitated reimbursement using nonlinear models. *Medical Care, 41*(6), 741-752.

* von Wyl, V., Beck, K. (2015). Do insurers respond to risk adjustment? A long-term, nationwide analysis from Switzerland. *European Journal of Health Economics*. DOI: 10.1007/s10198-015-0669-x.
* Vivas, D., Guadalajara, N., Barrachina, I., Trillo, J.–L, Usó, R., De-La-Poza, E. (2011). Explaining primary healthcare pharmacy expenses using classification of medications for chronic conditions. *Health Policy, 103*(1), 9-15.
* Ware, J.E. Jr., Kosinski, M., Keller, S.D. (1996). A 12-Item Short-Form Health Survey: Construction of Scales and Preliminary tests of Reliability and Validity. *Medical Care, 34*, 220-233.
* Weiner, J.P., Trish, E., Abrams, C., Lemke, K. (2012). Adjusting for risk selection in state health insurance exchanges will be critically important and feasible, but not easy. *Health Affairs, 31*(2), 306-315.
* Welch, W. (1985). Regression towards the mean in medical costs – implications for biased selection in HMOs. *Medical Care, 23*, 1234-1241.
* Wooldridge, J.M. (2003). *Introductory Econometrics. A modern Approach*. Second edition, South-Western, a division of Thomson Learning, United States of America.
* Wrobel, M.V., Doshi, J., Stuart, B.C., Briesacher, B. (2004). Predictability of prescription drug expenses for Medicare beneficiaries. *Health Care Financing Review, 25*(2), 37-46.
* Yu, W., Ellis, R.P., Ash, A.S. (2001). Risk Selection in the Massachusetts State Employee Health Insurance Program. *Health Care Management Sciences, 4*, 281-287.
* Yu, H., Dick, A.W. (2010). Risk-adjusted capitation rates for children: how useful are the survey-based measures? *Health Services Research, 45*(6, part 2), 1948-1962.
* Yuen, E.J., Louis, D.Z., Loreto, P.D., Gonnella, J.S. (2003). Modeling risk-adjusted capitation rates for Umbria, Italy. *European Journal of Health Economics, 4*(4), 304-312.
* Zhao, Y., Ellis, R.E., Ash, A.S., Calabrese, D., Ayanian, J.Z., Slaughter, J.P., Weyuker, L., Bowen, B. (2001). Measuring Population Health Risks Using Inpatient Diagnoses and Outpatient Pharmacy Data. *Health Services Research, 36*, 6, 180-193.
* Zhao, Y., Ash, A.S., Ellis, R.P., Ayanian, J.Z., Pope, G.C., Bowen, B., Weyuker, L. (2005). Predicting pharmacy expenses and other medical expenses using diagnoses and drug claims. *Medical Care, 43*(1), 34-43.

Samenvatting

AANLEIDING VAN HET ONDERZOEK

In verschillende landen wereldwijd – zoals België, Duitsland, Israël, Nederland, de Verenigde Staten, en Zwitserland, – is *gereguleerde concurrentie* tussen zorgverzekeraars geïntroduceerd. Concurrentie heeft als doel om kwalitatief goede en doelmatige zorg te stimuleren. Regulering waarborgt de publieke belangen als solidariteit en toegankelijkheid. Gereguleerde concurrentie kan alleen leiden tot kwalitatief goede en doelmatige zorg die voor iedere burger toegankelijk is als aan een aantal randvoorwaarden is voldaan. Eén cruciale randvoorwaarde is een adequaat risicovereveningssysteem.

Het risicovereveningssysteem beoogt zorgverzekeraars te compenseren voor voorspelbare verschillen in zorgkosten tussen verzekerden die gerelateerd zijn aan gezondheid. Bijvoorbeeld, er kan verwacht worden dat chronisch zieken hogere zorgkosten zullen hebben dan gezonde studenten. Zorgverzekeraars mogen de premie die zij vragen aan verzekerden niet differentiëren naar de voorspelbare zorgkosten van de verzekerde (verbod op premiedifferentiatie, Zorgverzekeringswet). Het gevolg van deze premieregulering is dat er financiële prikkels kunnen zijn om verzekerden te selecteren. Uiteindelijk kan risicoselectie de solidariteit, toegankelijkheid en kwaliteit van zorg in gevaar brengen. Het risicovereveningssysteem heeft als doel om de financiële prikkels tot risicoselectie weg te nemen. De mate waarin dit systeem dit beleidsdoel realiseert, wordt in belangrijke mate bepaald door de *voorspelkracht van het risicovereveningsmodel*. Het vereveningsmodel bevat een set van risicovereveningskenmerken die zorgkosten kunnen voorspellen, zoals leeftijd, geslacht en diagnose-informatie. Aan de hand van de vereveningskenmerken in het model worden voorafgaand aan elk kalenderjaar de risico-afhankelijke compensaties aan zorgverzekeraars berekend. Naar mate het risicovereveningsmodel beter in staat is om verwachte verschillen in zorgkosten tussen verzekerden te voorspellen, zullen de onder- en overcompensaties voor specifieke groepen in de populatie afnemen en zodoende zullen de financiële prikkels tot risicoselectie afnemen. Echter, indien het model niet in staat is om zorgverzekeraars adequaat te compenseren voor voorspelbare kostenverschillen tussen verzekerden zullen er financiële prikkels tot risicoselectie zijn. Een effectieve manier om deze financiële prikkels te reduceren is het verder verbeteren van de voorspelkracht van het model door het toevoegen van nieuwe vereveningskenmerken. Een andere manier om deze financiële prikkels te reduceren is om zorgverzekeraars achteraf te compenseren voor (een deel van de) werkelijke zorgkosten: de zogenoemde ex-post compensaties. Een nadeel van ex-post compensaties is echter dat deze de prikkels tot doelmatigheid verminderen.

DOEL VAN HET ONDERZOEK

Het *eerste deel* van dit proefschrift richt zich op het *evalueren* van de voorspelkracht van risicovereveningsmodellen. Het doel is om een aantal richtlijnen te formuleren voor het evalueren van deze modellen en het interpreteren van de uitkomsten van evaluatiestudies. Dit doel wordt gerealiseerd door: (i) het uitvoeren van een systematisch literatuuronderzoek naar evaluatiemethoden die gebruikt zijn om de voorspelkracht van risicovereveningsmodellen te meten. Dit houdt specifiek in dat gekeken is welke beoordelingsmaatstaven zijn gebruikt en op welke manieren deze maatstaven zijn toegepast; en (ii) het ontwikkelen van een evaluatiemethode om de potentiële afname in de financiële prikkels tot risicoselectie voor meerdere groepen in de populatie tezamen te meten. Deze evaluatiemethode geeft een algemeen beeld van de voorspelkracht van een risicovereveningsmodel voor verschillende groepen in de populatie. Deze methode is ontwikkeld om modellen te kunnen vergelijken op basis van de potentiële afname van de financiële prikkels tot risicoselectie om zodoende te bepalen welk model de voorkeur verdient in de praktijk.

Het *tweede deel* van het proefschrift richt zich op het *verbeteren* van de voorspelkracht van risicovereveningsmodellen. Het doel is het onderzoeken van de mate waarin drie nieuwe methoden de voorspelkracht van het Nederlandse model verder kunnen verbeteren. Dit gebeurt door het uitvoeren van diverse empirische analyses, waarin de bijdrage aan de voorspelkracht per methode is onderzocht.

DEEL I: HET EVALUEREN VAN DE VOORSPELKRACHT VAN RISICOVEREVENINGSMODELLEN

Om te bepalen in hoeverre een risicovereveningsmodel de prikkels tot risicoselectie reduceert en of een nieuw vereveningskenmerk voorspelkracht heeft, is het noodzakelijk om het model te evalueren. Desondanks is relatief weinig aandacht geschonken aan evaluatiemethoden en beoordelingsmaatstaven voor risicovereveningsmodellen. Het is echter belangrijk om goed inzicht te hebben in de kenmerken van beoordelingsmaatstaven en de manieren waarop deze maatstaven zijn toegepast, omdat verschillende maatstaven en/of verschillende toepassingen van dezelfde maatstaf kunnen resulteren in verschillende uitkomsten. Uiteindelijk zou dit kunnen leiden tot andere beslissingen over welk model de voorkeur verdient in de praktijk. In *hoofdstuk 2* is op basis van het systematisch literatuur onderzoek een classificatiesysteem voor beoordelingsmaatstaven ontwikkeld. Ook worden in dit hoofdstuk belangrijke kenmerken van deze maatstaven in combinatie met de toepassing(en) besproken. Op basis van dit onderzoek blijkt dat verschillende maatstaven zijn toegepast en dat eenzelfde maatstaf op verscheidene manieren is gebruikt om de voorspelkracht van risicovereveningsmodellen te meten. De voorkeur voor een bepaalde maatstaf hangt af van

de kenmerken van deze maatstaf in combinatie met de methode waarop deze maatstaf zal worden toegepast. Een belangrijke bevinding is dat slechts één specifieke evaluatiemethode geschikt is om te bepalen in hoeverre een risicovereveningsmodel prikkels tot risicoselectie reduceert: deze methode betreft het evalueren van de voorspelkracht van het model op het niveau van selectieve groepen in de populatie, bijvoorbeeld personen met een bepaalde chronische aandoening. Andere evaluatiemethoden die vaak worden toegepast – zoals het meten van de voorspelkracht op het niveau van alle individuen in de populatie, willekeurige groepen of verzekeraarsportefeuilles – bieden slechts beperkt inzicht in de financiële prikkels tot risicoselectie.

Om een algemeen beeld te verkrijgen van de mate waarin een risicovereveningsmodel de financiële prikkels tot risicoselectie reduceert is het nuttig om de voorspelkracht van het model voor meerdere selectieve groepen in de populatie te meten; bijvoorbeeld verschillende patiëntengroepen. Bij een dergelijke evaluatiestudie is het mogelijk dat tegenstrijdige uitkomsten verkregen worden met betrekking tot welk model de voorkeur verdient, omdat een model voor de ene groep de kosten beter kan voorspellen maar voor een andere groep slechter dan een ander model. *Hoofdstuk 3* ontwikkelt en toetst verschillende evaluatiemethoden die gebruikt kunnen worden om te bepalen in welke mate een model de financiële prikkels tot risicoselectie voor meerdere selectieve groepen tegelijkertijd reduceert. Elk van deze methoden corrigeert op een andere manier voor het dubbeltellen van personen die in meerdere groepen voorkomen, bijvoorbeeld personen met meerdere chronische aandoeningen. Vervolgens zijn deze evaluatiemethoden vergeleken met de meest eenvoudige methode die voor handen is, namelijk het optellen van de onder- en overcompensaties voor verschillende groepen zonder rekening te houden met dubbeltelling. Een belangrijke bevinding van dit onderzoek is dat voor de geanalyseerde gegevensbestanden de verschillende methoden (inclusief de meest eenvoudige) niet leiden tot verschillende conclusies met betrekking tot welk van de onderzochte modellen zorgt voor de grootste potentiële afname van de financiële prikkels tot risicoselectie en daarmee de voorkeur verdient.

Uit hoofdstuk 2 en 3 volgt een aantal richtlijnen voor het evalueren van de voorspelkracht van risicovereveningsmodellen, om zodoende te waken voor potentiële valkuilen bij het uitvoeren van een evaluatiestudie en het interpreteren van de uitkomsten van dergelijke studies. Bovendien beogen deze richtlijnen een aantal misvattingen in de literatuur over evaluatiemethoden en beoordelingsmaatstaven voor risicovereveningsmodellen weg te nemen. Een belangrijke misvatting is gerelateerd aan de methode die gebruikt wordt om selectieve groepen in de populatie te definiëren. Gezien het belang van het evalueren van de voorspelkracht op het niveau van selectieve groepen om inzicht te krijgen in de mate waarin een model financiële prikkels tot risicoselectie reduceert, is een aantal richtlijnen specifiek gericht op deze evaluatiemethode. Eén van deze richtlijnen geeft aan hoe groepen gedefinieerd dienen te worden: namelijk, de selectieve groepen mogen *niet* identiek zijn aan degenen die expliciet in het geëvalueerde model zijn opgenomen. De reden hiervoor is dat

een risicovereveningsmodel vanzelfsprekend de kosten goed voorspelt voor alle groepen die expliciet zijn opgenomen via de vereveningskenmerken, vanwege de schattingsmethode die momenteel gebruikt wordt.

Het is belangrijk om op te merken dat de richtlijnen niet stellen hoe risicovereveningsmodellen in elke situatie geëvalueerd dienen te worden en hoe de uitkomsten van evaluatiestudies in elke situatie geïnterpreteerd moeten worden. De keuze voor een bepaalde evaluatiemethode en de interpretatie van de uitkomsten van evaluatiestudies hangen namelijk af van: (i) de specifieke context waarin de evaluatiestudie is uitgevoerd; (ii), praktische beperkingen van deze studie, zoals de beschikbaarheid en kwaliteit van data; en (iii), de definitie van de vereveningskenmerken in de modellen die geëvalueerd zijn.

Daarnaast dient opgemerkt te worden dat de voorspelkracht slechts één van de criteria is om te bepalen welke risicovereveningsmodel gebruikt gaat worden in de praktijk, ook al is het een belangrijk en veelbesproken criterium. Andere evaluatiecriteria zijn de prikkels tot doelmatigheid, normatieve keuzes met betrekking tot risicofactoren waarvoor beleidsmakers compensatie wenselijk achten en de uitvoerbaarheid. Door andere afwegingen met betrekking tot deze criteria kunnen dezelfde uitkomsten van een evaluatiestudie resulteren in verschillende beslissingen over welk model gebruikt wordt. Dit betekent dan ook dat een model met een hogere voorspelkracht niet per definitie de voorkeur verdient boven een model met een lagere voorspelkracht, bijvoorbeeld vanwege de prikkels tot doelmatigheid die uitgaat bij het gebruik van een vereveningskenmerk gebaseerd op werkelijke zorgkosten in het voorgaande jaar. Bij de uiteindelijke beslissing over welk risicovereveningsmodel wordt gebruikt dienen alle evaluatiecriteria tezamen beoordeeld te worden.

DEEL II: HET VERBETEREN VAN DE VOORSPELKRACHT VAN RISICOVEREVENINGSMODELLEN

De voorspelkracht van het risicovereveningsmodel is in veel landen de afgelopen jaren sterk toegenomen door onder andere het ontwikkelen van vereveningskenmerken gebaseerd op diagnose-informatie op basis van het voorgaande jaar. Onderzoek toont echter aan dat deze uitgebreide modellen nog niet goed genoeg zijn, omdat nog forse onder- en overcompensaties voor specifieke groepen in de populatie bestaan. Om de financiële prikkels tot risicoselectie te reduceren is het van belang om de voorspelkracht van deze modellen verder te verbeteren. In de *hoofdstukken 4, 5 en 6* worden drie potentieel relevante nieuwe methoden ontwikkeld en getoetst. Deze methoden zijn gebaseerd op reeds beschikbare informatie uit de administratieve gegevens van zorgverzekeraars. Hierdoor zijn de kosten van dataverzameling minimaal en zouden de ontwikkelde modelverbeteringen tegen relatief lage administratieve kosten geïmplementeerd kunnen worden. Doordat de drie methoden op verschillende momenten zijn onderzocht zijn verschillende jaren van de administra-

tieve gegevensbestanden gebruikt. Het was hierdoor niet mogelijk om het onderzochte risicovereveningsmodel constant te houden en zodoende conclusies te trekken over welke methode tot de grootste verbetering in voorspelkracht leidt.

In *hoofdstuk 4* is onderzocht in welke mate de voorspelkracht van het Nederlandse risicovereveningsmodel van 2012 toeneemt door toevoeging van vereveningskenmerken die gebaseerd zijn op kosten- en/of diagnose-informatie uit drie voorgaande jaren. De analyse laat zien dat de voorspelkracht van dit model inderdaad verbeterd kan worden door toevoeging van dergelijke vereveningskenmerken. De R-kwadraat (R^2) op individuniveau kan met ongeveer 8 procentpunt toenemen en de 'Gewogen Gemiddelde Absolute Afwijking' (GGAA) kan met ongeveer € 125 afnemen, gegeven het Nederlandse model van 2012 met een R^2 van 28,54% en een GGAA van € 1.475. Echter, een dergelijk uitgebreid model blijkt nog steeds bepaalde specifieke groepen in de populatie onder of over te compenseren. Een voorbeeld van een groep die onder-gecompenseerd wordt zijn personen met één of meerdere zelf-gerapporteerde aandoeningen.

In *hoofdstuk 5* is onderzocht in welke mate de voorspelkracht van het Nederlandse risicovereveningsmodel van 2014 toeneemt door toevoeging van interactietermen tussen bestaande vereveningskenmerken. Een voorbeeld van interactietermen zijn de vier subgroepen die ontstaan door het combineren van de kenmerken gezond/ongezond en 65-min/65-plus. Interactietermen zijn momenteel beperkt toegepast, terwijl zij ervoor zouden kunnen zorgen dat de kosten beter voorspeld worden voor specifieke groepen in de populatie, zoals personen met co-morbiditeit. Omdat het grote aantal bestaande vereveningskenmerken in combinatie met de grootte van de databestanden in theorie de definitie van een zeer groot aantal interactietermen mogelijk maakt, is gebruik gemaakt van 'regression tree modelling'. Deze analysetechniek identificeert alle interactietermen die een statistisch significante bijdrage leveren aan het verklaren van de verschillen in zorgkosten tussen verzekerden die nog niet verklaard worden door de vereveningskenmerken in het Nederlandse risicovereveningsmodel van 2014. Vervolgens zijn deze interactietermen aan dit model toegevoegd om te bepalen in hoeverre de voorspelkracht toeneemt. De analyse toont dat interactietermen inderdaad kunnen zorgen voor een verbetering van de voorspelkracht. De R^2 op individuniveau kan met ongeveer 0,08 tot 1,78 procentpunt toenemen en de GGAA met ongeveer € 0 tot € 9 afnemen (afhankelijk van de specificatie van het 'regression tree model'), gegeven het Nederlandse risicovereveningsmodel van 2014 met een R^2 van 25,56% en een GGAA van € 1.569. Echter, een dergelijk uitgebreid model blijkt nog steeds bepaalde specifieke groepen in de populatie onder te compenseren, zoals personen met één of meerdere zelf-gerapporteerde aandoeningen.

In *hoofdstuk 6* is onderzocht in welke mate de voorspelkracht van het Nederlandse risicovereveningsmodel van 2013 toeneemt door toevoeging van één vereveningskenmerk of meerdere interactietermen die gebaseerd zijn op informatie over de onder- en overcompensaties in drie voorgaande jaren. Om dit vereveningskenmerk of de interactietermen te

definiëren is het noodzakelijk om eerst vast te stellen of er een groep bestaat die structureel ondergecompenseerd wordt over drie voorgaande jaren. Om deze reden is onderzocht of een dergelijke groep bestaat en wat de kostenpatronen en risicokenmerken van de personen in deze groep zijn. De analyse bevestigt dat het Nederlandse risicovereveningsmodel van 2013 structureel een groep ondercompenseert over drie voorgaande jaren: deze groep omvat ongeveer 1% tot 4% van de populatie, afhankelijk van de definitie die gebruikt wordt. De personen in deze groep hebben in het algemeen bovengemiddelde zorgkosten in elk jaar en hebben veelal één of meerdere zelf-gerapporteerde langdurige aandoeningen. Het toevoegen van één vereveningskenmerk of meerdere interactietermen voor specifiek de groep personen die structureel ondergecompenseerd worden, verbetert de voorspelkracht van het Nederlandse risicovereveningsmodel van 2013. Toevoeging van één vereveningskenmerk kan de R^2 op individuniveau van dit model met ongeveer 0,05 tot 0,10 procentpunt laten toenemen en de GGAA met ongeveer € 1 tot € 9 laten afnemen (het Nederlandse risicovereveningsmodel van 2013: R^2 is 24,22% en GGAA is € 1.570). Toevoeging van interactietermen kan de R^2 op individuniveau van dit model met ongeveer 1,51 tot 1,64 procentpunt laten toenemen en de GGAA met € 5 tot € 13 laten afnemen. Echter een dergelijk uitgebreid model blijkt wederom nog steeds specifieke groepen in de populatie onder en over te compenseren.

De bevindingen van de hoofdstukken 4, 5, en 6 tonen aan dat de voorspelkracht van risicoverveningsmodellen verbeterd kan worden door toevoeging van vereveningskenmerken op basis van kosten en/of diagnose-informatie uit drie voorgaande jaren, of door interactietermen tussen bestaande vereveningskenmerken, of door een risicovereveningkenmerk of meerdere interactietermen op basis van informatie over de onder- en overcompensaties in drie voorgaande jaren. Echter, na toepassing van elk van deze drie nieuwe methoden afzonderlijk blijken nog steeds onder- en overcompensaties voor specifieke groepen in de populatie te bestaan. Dit leidt tot de conclusie dat elk van de drie onderzochte methoden de financiële prikkels tot risicoselectie vermindert maar niet wegneemt.

BELEIDSAANBEVELINGEN

Het onderzoek in dit proefschrift leidt tot een zestal beleidsaanbevelingen. De eerste aanbeveling is om de voorspelkracht van risicovereveningsmodellen te evalueren op specifieke groepen in de populatie. De reden hiervoor is dat dit de enige evaluatiemethode is die goed inzicht kan geven in de mate waarin het risicovereveningsmodel financiële prikkels tot risicoselectie reduceert. Voor het uitvoeren van dergelijke evaluatiestudies en het interpreteren van de uitkomsten formuleert dit proefschrift een aantal richtlijnen. Eén belangrijke richtlijn is dat de groepen waarop een risicovereveningsmodel wordt geëvalueerd *niet* identiek

mogen zijn aan de groepen die expliciet zijn opgenomen via de vereveningskenmerken in dat model (gegeven de huidige schattingsmethode).

Een tweede beleidsaanbeveling is dat – indien kwalitatief goede informatie niet beschikbaar is om groepsniveau-evaluaties uit te voeren – moet worden geïnvesteerd in het verkrijgen van dergelijke informatie, bijvoorbeeld door het uitzetten van gezondheidsenquêtes. Bij voorkeur wordt deze informatie op jaarlijkse basis verzameld, zodat de financiële prikkels tot risicoselectie onder het risicovereveningsmodel dat gebruikt wordt over de tijd gemonitord kunnen worden voor dezelfde selectieve groepen in de populatie.

Een derde aanbeveling is dat de voorspelkracht van risicovereveningsmodellen verder verbeterd kan worden door het toepassen van elk van de drie onderzochte methoden afzonderlijk. Van deze methoden is het gebruik van vereveningskenmerken gebaseerd op diagnose-informatie van drie voorgaande jaren een eerste aantrekkelijke optie, omdat vereveningskenmerken gebaseerd op diagnose-informatie van één voorafgaand jaar al gebruikt worden in het huidige Nederlandse risicovereveningsmodel. Hierbij dient opgemerkt te worden dat het Nederlandse risicovereveningsmodel tijdens het uitvoeren van dit onderzoek is uitgebreid met nieuwe vereveningskenmerken en dat de zorgkosten die meegenomen worden in het model ook uitgebreid zijn. Om deze reden verdient het aandacht om de drie onderzochte methoden opnieuw door te rekenen op basis van het huidige risicovereveningsmodel met de huidige kostendefinities. Bij een dergelijk vervolgonderzoek dient ook rekening gehouden te worden met andere evaluatiecriteria dan alleen de voorspelkracht, zoals de prikkels tot doelmatigheid.

Een vierde aanbeveling is dat vervolgonderzoek naar het verder verbeteren van de risicovereveningsmodellen noodzakelijk is. Dit proefschrift laat zien dat elk van de drie onderzochte nieuwe methoden afzonderlijk de financiële prikkels tot risicoselectie niet kunnen wegnemen. Mogelijk zouden de drie methoden gecombineerd kunnen worden. Indien het combineren van de drie methoden nog steeds selectieve groepen onder- en overcompenseren – wat niet onwaarschijnlijk is – is het aanbevolen om op zoek te gaan andere informatiebronnen dan de administratieve gegevensbestanden van zorgverzekeraars, bijvoorbeeld het verzamelen van specifieke informatie over zeldzame aandoeningen, personen in de laatste levensjaren (i.e. mortaliteit) en zwangerschap.

Een vijfde aanbeveling is om te overwegen om alternatieve effectieve strategieën, zoals ex-post compensaties, toe te passen om de financiële prikkels tot risicoselectie voor gericht specifieke groepen in de populatie te reduceren zolang het risicovereveningsmodel niet in staat is om de kosten voor deze groepen adequaat te voorspellen. Zo zouden bijvoorbeeld zorgverzekeraars deels risicodragend kunnen worden voor de kosten van ongeveer 1% - 4% van de verzekerden in de populatie die structureel ondergecompenseerd worden over drie voorgaande jaren en volledig risicodragend voor de kosten van de overige verzekerden in de populatie. Dergelijke ex-post compensaties kunnen tijdelijk ingezet worden totdat adequate vereveningskenmerken zijn ontwikkeld.

Tot slot, een zesde aanbeveling is om het evalueren en het verbeteren van het risicovereveningsmodel effectief te coördineren en onderdeel te maken van een continue beleidscyclus. Door het evalueren en verbeteren van het model elkaar te laten opvolgen zouden substantiële verbeteringen in het reduceren van financiële prikkels tot risicoselectie gerealiseerd kunnen worden. Het evalueren van het vereveningsmodel kan namelijk inzicht geven in de financiële prikkels tot risicoselectie en daarmee voor welke groepen verbetering is vereist. Deze informatie kan vervolgens ingezet worden om gerichte modelverbeteringen door te voeren. Het evalueren van het aangepaste vereveningsmodel geeft weer inzicht in de mate waarin de financiële prikkels tot risicoselectie zijn gereduceerd en voor welke groepen nu verbetering is vereist. Als onderdeel van de beleidscyclus kunnen acties die wijzen op risicoselectie gemonitord worden. Indien nodig kan deze informatie ingezet worden om de voorspelkracht van het vereveningsmodel verder te verbeteren om zodoende de financiële prikkels voor gerichte groepen te reduceren.

AANBEVELINGEN VOOR VERVOLGONDERZOEK

De bevinding van dit onderzoek dat financiële prikkels tot risicoselectie niet geëlimineerd kunnen worden door afzonderlijke toepassing van de drie onderzochte methoden onderschrijft het belang van vervolgonderzoek naar het verder verbeteren van de voorspelkracht van risicovereveningsmodellen. Het is relevant om in dit vervolgonderzoek specifiek aandacht te geven aan verzekerden die structureel onder-gecompenseerd worden. Het is namelijk goed mogelijk dat deze groep (voor een deel) bestaat uit verzekerden met (extreem) hoge zorgkosten voor wie het moeilijk is om de zorgkosten adequaat te voorspellen, bijvoorbeeld personen met zeldzame aandoeningen, personen in de laatste levensjaren en zwangere vrouwen. Gedetailleerde informatie over de kosten en risicokenmerken van de groep(en) die structureel onder-gecompenseerd worden kan waardevolle inzichten opleveren over mogelijke modelverbeteringen en de noodzaak van alternatieve maatregelen, zoals ex-post compensaties. Bij dergelijk vervolgonderzoek is het belangrijk om rekening te houden met hoe de verschillende evaluatiecriteria, zoals prikkels tot risicoselectie en doelmatigheid, afgewogen worden. Het beoordelingskader van de verschillende evaluatiecriteria geeft namelijk de mogelijkheden dan wel grenzen aan van de mate waarin de voorspelkracht van het risicovereveningsmodel verder verbeterd kan worden en de mogelijkheden tot het (tijdelijk) inzetten van ex-post compensatie maatregelen.

Dankwoord

Het moment is aangebroken om een speciaal woord te richten aan de personen die in welke vorm dan ook hebben bijgedragen aan dit proefschrift. Bij het schrijven van deze woorden flitsen de afgelopen vijf jaren in mijn hoofd voorbij… Ik kom al snel tot de conclusie dat ik een groot aantal personen heb mogen ontmoeten, met veel personen heb mogen samenwerken en veel leuke en mooie momenten heb meegemaakt. Dit gegeven maakt het dan ook moeilijk om de juiste woorden te vinden om een ieder op een manier te bedanken die recht doet aan de tijd en inzet, dan wel de leerzame en wijze lessen om mezelf te verbeteren, dan wel de leuke momenten die we hebben gehad. Bovendien realiseer ik mij nu dat het er simpelweg gewoon te veel zijn om hier een woord te richten aan iedereen. Ik zal daarom helaas hier mijn woord richten aan een beperkt aantal personen, ook al betekent dit niet dat de bijdrage van anderen minder belangrijk zijn geweest voor mij.

In de eerste plaats wil ik hier graag drie personen bedanken die centraal stonden tijdens het schrijven van dit proefschrift: mijn promotor Wynand van de Ven en de copromotoren Richard van Kleef en René van Vliet. Het is van onschatbare waarde om een proefschrift over risicoverevening te schrijven onder begeleiding van deze drie personen. Het was een unieke ervaring om naast het doen van wetenschappelijk onderzoek naar risicoverevening ook daadwerkelijk in de praktijk bezig te zijn met de risicoverevening. Deze ervaring draag ik voor altijd met mij mee. Ik heb de samenwerking met jullie als zeer prettig ervaren.

Wynand, aan jou ben ik veel dank verschuldigd. Tijdens het sollicitatiegesprek vroeg je mij of ik nog steeds zoveel uren per week besteedde aan het wielrennen omdat dat echt niet meer zou kunnen en of ik wel zeker wist dat ik wilde promoveren, omdat het toch echt iets anders is dan advieswerk. Veel dank dat je mij toen de kans hebt gegeven om een proefschrift te schrijven. Ik wil je ook bedanken voor de tijd en energie die je genomen hebt om tussentijdse stukken te voorzien van commentaar en suggesties.

Richard, ook aan jou ben ik veel dank verschuldigd. Ondanks je drukke agenda was het altijd mogelijk om een afspraak in te plannen. Ook veel dank voor jouw tijd en energie om de tussentijdse stukken van commentaar en suggesties te voorzien. Het was daarnaast heel fijn dat je vaak zo even langs kon lopen om iets te vragen.

René, last but not least, ook aan jou ben ik veel dank verschuldigd. Jouw kennis over risicoverevening en statistiek is buitengewoon. Ik heb daar veel van kunnen leren en daar ben ik je voor altijd dankbaar voor. Ook dank voor jouw tijd en inzet om tussentijdse stukken van commentaar en suggesties te voorzien. Als laatste, vind ik het bewonderingswaardig hoe jij het belangrijke onderzoeksproject voor het berekenen van de normbedragen begeleidt.

Ook ben ik veel dank verschuldigd aan prof. dr. E.K.A. van Doorslaer, prof. dr. E. Schokkaert en prof. dr. J. Boone (beoordelingscommissie) en aan prof. dr. J.J. Polder en dr. H.A. Keuzekamp (promotiecommissie) voor het lezen en beoordelen van mijn proefschrift en het opponeren bij de verdediging.

Daarnaast gaat ook mijn dank uit naar een aantal andere personen die naast Wynand, Richard en René, in welke vorm dan ook, hebben bijgedragen aan de inhoud van dit proefschrift. Zonder anderen te kort te doen denk ik aan de leden van de Risk Adjustment Network en de redacteuren en reviewers van de tijdschriften waar verschillende hoofdstukken van dit proefschrift zijn ingediend.

Een groot deel van dit proefschrift bestaat uit empirische analyses die niet uitgevoerd hadden kunnen worden zonder toestemming voor het gebruik van de administratieve gegevensbestanden. Ik wil alle leden van de Begeleidingscommissie bedanken voor het beschikbaar stellen van deze bestanden voor wetenschappelijk onderzoek. Ook wil ik het Centraal Bureau voor de Statistiek bedanken voor het versleutelen van het identificatiekenmerk in de administratieve gegevensbestanden om zodoende deze bestanden te kunnen koppelen aan de gezondheidsenquêtes.

Ook wil ik hier graag mijn collega's, in het bijzonder die van de sectie ZKV: Anne-Fleur, Danielle (C.), Danielle (D.), Edith, Erik, Frank, Kayleigh, Marco, Rudy, Stephanie en Trea, bedanken voor de fijne werkomgeving. Van deze collega's wil ik Danielle (C.), Danielle (D.) en Kayleigh extra bedanken. Jullie waren fijne kamergenoten en dat heeft zeker bijgedragen aan leuke werkdagen. Dank voor jullie interesse en gezelligheid. Ook wil ik nog Frank bedanken voor de prettige samenwerking bij het uitvoeren van het onderzoeksproject voor het berekenen van de normbedragen.

Als laatste, wil ik het woord richten aan een kleine groep personen die heel erg dicht bij mij staan. Jullie allen zorgen ervoor dat het leven een groot plezier is. We delen geluk en verdriet samen en jullie betekenen ontzettend veel voor mij. Het is een geweldige gedachte om na elke werkdag en werkweek op weg naar huis te weten dat ik in jullie gezelschap kan zijn. Ik hoop dan ook nog met jullie van het leven te genieten en nog vele mooie momenten met jullie te delen.

Curriculum Vitae

Suzanne H.C.M. van Veen

Date and place of birth: 3 October 1987, Delft (the Netherlands)

PROFESSIONAL WORK EXPERIENCES

* Doctoral candidate / scientific researcher at the institute of Health Policy and Management (iBMG), Erasmus University Rotterdam (EUR), in the Netherlands — 2011-2015
  - PhD thesis: "Evaluating and Improving the Predictive Performance of Risk Equalization Models in Health Insurance Markets"
  - Member of the research-team that worked on calculating the actual risk-equalization model in the Netherlands (2013-2015)
  - Teaching courses at the iBMG, EUR: statistics and multivariate analyses
  - (International) Conferences and presentations
  - Post-academic course on risk equalization to relevant stakeholders in the Dutch healthcare market (2014-2105)
* Consultant, Strategies in Regulated Markets, the Hague — 2010
* Internship, Strategies in Regulated Markets, the Hague (4 months) — 2010
* Internship, Fortis Health Care, Fortis Bank, Rotterdam (4 months) — 2009
* Internship, Medisch Centrum Haaglanden, the Hague (4 months) — 2009

EDUCATION

* M.Sc. in "Health Economics Policy and Law", specialization in "Health Economics", Erasmus University Rotterdam, the Netherlands — 2009-2010
* B.Sc. in "Algemene gezondheidswetenschappen", Erasmus University Rotterdam, the Netherlands — 2006-2009

SCIENTIFIC PEER-REVIEWED PUBLICATIONS

Published or under review

* van Veen, S.H.C.M., van Kleef, R.C., van de Ven, W.P.M.M., van Vliet, R.C.J.A. (2015). Improving the Prediction Model used in Risk Equalization: Cost and Diagnostic Information from Multiple Prior Years. *European Journal of Health Economics, 16*(2), 201-218.

* 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 Risk equalization Models. *Medical Care Research and Review, 72*(2), 220-243
* van Veen, S.H.C.M., van Kleef, R.C., van de Ven, W.P.M.M., van Vliet, R.C.J.A. (2015). Evaluating Risk Equalization Models by estimating the Potential Selection Profits. *Submitted for publication.*
* van Veen, S.H.C.M., van Kleef, R.C., van de Ven, W.P.M.M., van Vliet, R.C.J.A. (2015). Exploring the Predictive Power of Interaction Terms in a Sophisticated Risk Equalization Model using Regression Trees. *Submitted for publication.*
* van Veen, S.H.C.M., van Kleef, R.C., van de Ven, W.P.M.M., van Vliet, R.C.J.A. (2015). Exploring Persistent Under-compensations under a Morbidity-based Risk Equalization Model: Evidence from the Netherlands. *Submitted for publication.*

DUTCH NON-PEER REVIEWED PUBLICATIONS

* van Veen, S.H.C.M. (2015). Het belang van adequate risicoverevening. *De Actuaris*, thema Zorg en AWBZ. Koninklijk Actuarieel Genootschap, 22(3):14-16.

CONTRIBUTIONS TO DUTCH POLICY RESEARCH PROJECTS

* iBMG projectteam. (2015). Onderzoek risicoverevening 2016: Gegevensfase. WOR 747. Onderzoek voor het Ministerie van Volksgezondheid, Welzijn en Sport. iBMG, Erasmus Universiteit Rotterdam, te Rotterdam.
* iBMG projectteam. (2015). Onderzoek risicoverevening 2016: Overall Toets. WOR 748. Onderzoek voor het Ministerie van Volksgezondheid, Welzijn en Sport. iBMG, Erasmus Universiteit Rotterdam, te Rotterdam.
* iBMG projectteam. (2015). Onderzoek risicoverevening 2016: Berekening normbedragen. WOR 749. Onderzoek voor het Ministerie van Volksgezondheid, Welzijn en Sport. iBMG, Erasmus Universiteit Rotterdam, te Rotterdam.
* iBMG projectteam. (2014). Onderzoek risicoverevening 2015: Gegevensfase. WOR 709. Onderzoek voor het Ministerie van Volksgezondheid, Welzijn en Sport. iBMG, Erasmus Universiteit Rotterdam, te Rotterdam.
* iBMG projectteam. (2014). Onderzoek risicoverevening 2015: Overall Toets. WOR 710. Onderzoek voor het Ministerie van Volksgezondheid, Welzijn en Sport. iBMG, Erasmus Universiteit Rotterdam, te Rotterdam.

* iBMG projectteam. (2014). Onderzoek risicoverevening 2015: Berekening normbedragen. WOR 711. Onderzoek voor het Ministerie van Volksgezondheid, Welzijn en Sport. iBMG, Erasmus Universiteit Rotterdam, te Rotterdam.
* iBMG projectteam. (2014). Vervolgonderzoek risicoverevening 2015: voorspellende waarde van het gebruik van fysiotherapie voor de zorgkosten. WOR 712. Onderzoek voor het Ministerie van Volksgezondheid, Welzijn en Sport. iBMG, Erasmus Universiteit Rotterdam, te Rotterdam.
* iBMG projectteam. (2013). Onderzoek risicoverevening 2014: Gegevensfase. WOR 648. Onderzoek voor het Ministerie van Volksgezondheid, Welzijn en Sport. iBMG, Erasmus Universiteit Rotterdam, te Rotterdam.
* iBMG projectteam. (2013). Onderzoek risicoverevening 2014: Overall Toets. WOR 649. Onderzoek voor het Ministerie van Volksgezondheid, Welzijn en Sport. iBMG, Erasmus Universiteit Rotterdam, te Rotterdam.
* iBMG projectteam. (2013). Onderzoek risicoverevening 2014: Berekening normbedragen. WOR 658. Onderzoek voor het Ministerie van Volksgezondheid, Welzijn en Sport. iBMG, Erasmus Universiteit Rotterdam, te Rotterdam.
* iBMG projectteam. (2013). Vervolgonderzoek risicoverevening 2014: Het somatische vereveningsmodel 2014 uitgebreid met geriatrische revalidatiezorg. Onderzoek voor het Ministerie van Volksgezondheid, Welzijn en Sport. iBMG, Erasmus Universiteit Rotterdam, te Rotterdam.
* iBMG projectteam. (2013). Vervolgonderzoek risicoverevening 2014: Het somatische vereveningsmodel 2014 inclusief geriatrische revalidatiezorg uitgebreid met verpleging en verzorging. Onderzoek voor het Ministerie van Volksgezondheid, Welzijn en Sport. iBMG, Erasmus Universiteit Rotterdam, te Rotterdam.
* iBMG projectteam. (2013). Vervolgonderzoek risicoverevening 2014: Het GGZ-model 2014 uitgebreid met langdurige geestelijke gezondheidszorg. Onderzoek voor het Ministerie van Volksgezondheid, Welzijn en Sport. iBMG, Erasmus Universiteit Rotterdam, te Rotterdam.

INTERNATIONAL CONFERENCES

*	Risk Adjustment Network, Solothurn, Switzerland	2015
*	Risk Adjustment Network, Delft, the Netherlands	2014
*	International Health Economics Association/European Conference Health Economics, Dublin, Ireland	2014
*	Risk Adjustment Network, Tel Aviv, Israel	2013
*	Risk Adjustment Network, Lucerne, Switzerland	2012
*	European Conference Health Economics, Zurich, Switzerland	2012
*	International Health Economics Association, Toronto, Canada	2011

PRESENTATIONS AT INTERNATIONAL CONFERENCES

Paper presentations and/or discussant

*	Risk Adjustment Network conference, Solothurn, Switzerland	2015
*	Risk Adjustment Network conference, Delft, the Netherlands	2014
*	International Health Economics Association/European Conference Health Economics, Dublin, Ireland	2014
*	Risk Adjustment Network conference, Tel Aviv, Israel	2013
*	Risk Adjustment Network conference, Lucerne, Switzerland	2012
*	European Conference Health Economics, Zurich, Switzerland	2012

PRESENTATIONS AT NATIONAL CONFERENCES, COURSES AND DUTCH GOVERNMENT AGENCIES

*	Post-academic course "Risicoverevening: wat, waarom en waar staan we?", Institute of Health Policy and Management, Erasmus University Rotterdam	2014-2015
*	Dutch Ministry of Health, Welfare, and Sports, the Hague	2014
*	Koninklijk Actuarieel Genootschap, Utrecht	2014
*	Dutch Ministry of Health, Welfare, and Sports, the Hague	2013
*	LoLAHESG, Vereniging voor Gezondheidseconomie, Nunspeet	2013

ACADEMIC TRAINING AND COMPETENCE COURSES

*	Training "Professionele adviesvaardigheden" [In Dutch]	2015
*	Workshop "Personal Branding"	2014
*	E-learning "Applied Analytics using SAS® Enterprise Miner"	2013-2014
*	Workshop "Mediatraining" [In Dutch]	2013
*	Training "Academic Writing"	2012
*	Training "English Speaking"	2012
*	Personal training, "Presentation skills" [In Dutch]	2012
*	SAS® Base Programmer 1 and 2 [in Dutch]	2011
*	Training "Ready within four years" [in Dutch]	2011
*	Course "Modeling Health Care Costs and Counts", Toronto, Canada	2011

CERTIFICATES

*	Certified SAS® Base Programmer	2013 - present

TEACHING CREDENTIALS / TRAINING

*	"Geven van Onderwijs 2" [in Dutch]	2014
*	"Schijfopdrachten construeren, begeleiden en beoordelen" [in Dutch]	2012
		2012
*	"Module Toetsing 1" [in Dutch]	2012
*	"Geven van Onderwijs 1" [in Dutch]	2012
*	"Werkbijeenkomst bachelor 1" [in Dutch]	

TEACHING EXPERIENCES

*	"Statistics A", tutor and lecturer in pre-master program	2012-2015
*	"Statistics B", tutor and lecturer in pre-master program	2012-2015
*	"Statistics", tutor in bachelor program	2011-2013
*	"SPSS computer-sessions", tutor in pre-master and bachelor program	2011-2015
*	"Multivariate Analyses", tutor in bachelor program	2011-2015
*	"Quantitative Research", tutor in the pre-master	2011-2014
*	Co-supervisor of bachelor and masters theses	2011-2014

PROFESSIONAL TOPSPORT EXPERIENCES

*	Professional cyclist in international women cycling. A member in two UCI teams: "Buitenpoort-Flexpoint" (2005 – 2008) and "Leontien.nl" (2009), and several international races and championships for the Dutch National cycling road team.	2005 - 2009

WORK EXPERIENCES IN COMMITTEES AND ORGANIZING EVENTS

*	Member of the Risk Adjustment Network	2013- present
*	Organizer of the post-academic course [in Dutch] "Risicover-evening: wat, waarom en waar staan we?", at the iBMG, Erasmus University Rotterdam	2014-2015
*	Co-organizer of the Risk Adjustment Network Conference, Delft, the Netherlands	2014
*	President of the Erasmus PhD Association Rotterdam Board and university-representative for the PhD National Network Netherlands	2011-2013

About the author

Suzanne van Veen (1987) was a doctoral candidate and a scientific researcher at the Institute of Health Policy and Management (iBMG), Erasmus University Rotterdam. From 2006 to 2010 she studied Health Policy and Management at the Erasmus University Rotterdam. During her study period, she was a professional cyclist and competed at several international stage races, world trophy races and championships. In 2010, she obtained her master's degree in Health Economics, Policy and Law, with a specialization in Health Economics. During 2010, she also worked as a consultant at 'Strategies in Regulated Markets' (SiRM).

In 2011, she started working as a doctoral candidate and researcher at iBMG. From 2011 to 2015 she worked on her PhD dissertation and published her research in peer-reviewed scientific journals, including European Journal of Health Economics (Q1) and Medical Care Research and Review (Q1). In addition, she presented her work at several international conferences, including the international Health Economics Association in 2013 (iHEA), the European Conference on Health Economics in 2012 (ECHE), and the yearly conferences of the Risk Adjustment Network (RAN) during 2012 to 2015. Since 2013, she is an official member of the RAN.

In addition to the research included in her dissertation, Suzanne worked as a member of the research-team that worked on several projects on risk equalization in order to calculate the actual risk equalization model in the Netherlands.

As a teacher, she was involved in the bachelor and pre-master program of the iBMG at the Erasmus University Rotterdam as a tutor for the course "Multivariate Analyses" and as a tutor and lecturer for the course "Statistics". She also provided several post-academic courses on risk equalization to policymakers, health insurers, healthcare providers, advisors and other stakeholders in the Dutch healthcare market.

Several countries world-wide, including Belgium, Germany, Israel, the Netherlands, Switzerland, and the U.S., use a risk equalization (RE) model to provide risk-adjusted payments to health insurers. The goal of RE is to mitigate financial incentives for risk selection and thereby to achieve a level playing field for health insurers. The extent to which an RE-model achieves this goal depends on the predictive performance of this model. In contrast to the vast amount of literature paying attention to improving the predictive performance of RE-models, evaluating model's performance has been understudied. The first part of this thesis formulates general principles on how to evaluate model's performance. These principles may assist researchers and policymakers by performing empirical evaluations and interpreting the results of these evaluations for decision-making. Despite RE-models have been developed over the past decades, a critical question for policymakers in all countries with RE is still how to further improve model's predictive performance. The second part of this thesis examines three potentially relevant methods to improve the predictive performance of sophisticated morbidity-based RE-models.

Suzanne H.C.M. van Veen (1987) was a doctoral candidate and research fellow at the institute of Health Policy and Management, Erasmus University Rotterdam. She published her work in various peer-reviewed scientific journals and presented her work at several international and national conferences. In addition to this, she worked on several projects on risk equalizationin the Netherlands. She also provided post-academic courses on risk equalization to Dutch policymakers, insurers, providers and advisors.

ISBN: 978-94-6169-781-3