Testing the effectiveness of risk equalization models in health insurance

A new method and its application

ISBN 978-90-8559-152-8

© P.J.A. Stam, 2007

Cover photo credit: Marc Dietrich, Freiburg im Breisgau, Baden-Wuerttemberg, Germany

Printed by: Optima Grafische Communicatie, Rotterdam, The Netherlands

Testing the effectiveness of risk equalization models in health insurance

A new method and its application

Het toetsen van de effectiviteit van risicovereveningsmodellen voor zorgverzekeringen

Een nieuwe methode en haar toepassing

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Erasmus Universiteit Rotterdam op gezag van de Rector Magnificus Prof.dr. S.W.J. Lamberts en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op woensdag 17 oktober 2007 om 9.45 uur.

door

Pieter Johannes Adrianus Stam
geboren te Schoonhoven

PROMOTIECOMMISSIE

Promotor:
Prof.dr. W.P.M.M. van de Ven

Overige leden:
Prof.dr. E.K.A. van Doorslaer
Prof.dr. N.S. Klazinga
Prof.dr. E. Schokkaert

Copromotor:
Dr. R.C.J.A. van Vliet

To my mother

CONTENTS

1. Introduction

1.1 The purpose of this study 12
1.2 Risk equalization in the Netherlands 15
1.3 Research questions 19
1.4 Outline 20

Appendix chapter 1
A1.1 Company names of Dutch sickness funds 23
A1.2 The Dutch REF models for sickness funds 1991 – 2005 24

2. Theoretical framework and methods

2.1 A test for the effectiveness of risk-adjusted premium subsidies 30
2.1.1 Risk-adjusted premium subsidies in competitive health insurance markets 30
2.1.2 Causes for incomplete and imperfect REF adjusters 33
2.1.3 Methodological solutions for imperfect REF adjusters 37
2.1.4 Methodological solutions for incomplete REF adjusters 39
2.1.5 Additional regulations for improving the subsidies 40
2.2 Risk adjusters in the literature 43
2.3 Methods 50
2.3.1 REF predicted costs as an approximation to normative costs 50
2.3.2 The determination of normative costs 52
2.3.3 Aligning the REF weights with normative costs 56
2.3.4 Testing alternative specifications of the REF equation 60
2.3.5 Additional regulations for improving the subsidies 61
2.4 Conclusions 62

3. Data

3.1 Agis administrative data 1997-2002 66
3.2 Agis Health Survey 2001 67
3.2.1 Questionnaire design 68
3.2.2 Statistical representativeness 75
3.2.3 Selection of eligible cases for the analysis sample 80
3.3 External data sources 82
3.4 Conclusions 84

Appendix chapter 3
A3.1 Mode of data collection 86
A3.2 Predictive power of Z&Z survey variables 90
A3.3 Data collection process 91
A3.4 Description of Agis Health Survey 2001 items 98
A3.5 Response and nonresponse analysis 100
A3.6 Summary of administrative, survey and external variables 110

4. Health status measurement
4.1 Construction of eight SF-36 health scales 114
4.1.1 Completeness 119
4.1.2 Reliability 120
4.1.3 Validity 122
4.2 Conclusions 132

Appendix chapter 4
A4.1 The SF-36 item responses 137
A4.2 Response Consistency Index 140
A4.3 The SF-36 Physical and Mental Component Scales 145

5. A derivation of normative costs
5.1 A statistical description of the S-type adjusters 148
5.2 Estimation of the normative equation 156
5.3 Conclusions 162

Appendix chapter 5
A5.1 Diagnosic Cost Group classification 164

6. Testing the REF equation for effectiveness
6.1 A comparison of REF predicted costs and normative costs 168
6.1.1 Comparing subgroups defined by the S-type adjusters 170
6.1.2 Comparing subgroups defined by the REF adjusters 176
6.2 An adjustment of the REF weights by the omitted variables approach 179
6.3 A normative adjustment of the REF weights 187
6.4 Conclusions 192

Appendix chapter 6
A6.1 A normative test of the pre-2004 Dutch REF equations 195
A6.2 A supplement to Section 6.2 and Section 6.3 198

7. Testing alternative REF model specifications for effectiveness

7.1 Adding new risk adjusters to the REF equation 204
7.2 Ex-post risk sharing as a supplement to the REF equation 216
7.3 The functional form and error distribution of the REF equation 219
7.4 Conclusions 232

Appendix chapter 7
A7.1 A supplement to Section 7.1 235
A7.2 A supplement to Section 7.2 236
A7.3 A supplement to Section 7.3 238
A7.4 Normative test results of Section 7.3 after the exclusion of outliers 239

8. Premium rate restrictions to improve affordability

8.1 The tradeoff between affordability and selection 244
8.2 Risk rating across Dutch provinces 250
8.3 Conclusions 252

Appendix chapter 8
A8.1 Results for socio-economic subgroups 254

9. Conclusions and discussion

9.1 Conclusions 260
9.2 Discussion 267

Glossary 275
Abbreviations 279
References 281

Samenvatting

Het toetsen van de effectiviteit van risicovereveningsmodellen voor zorgverzekeringen 291

Curriculum Vitae 313
Acknowledgments 315

Chapter 1

Introduction

1.1 The purpose of this study

In several European countries (Belgium, Germany, Switzerland and The Netherlands), competition among health insurers is used to stimulate efficiency and responsiveness to consumers' preferences in the health care sector (Van de Ven et al. 2003). The ultimate goal is to stimulate health insurance companies to act as prudent purchasers or providers of care for their members. In an unregulated competitive health insurance market, however, insurers will risk rate their premiums according to an individual consumer's risk profile: sick people will pay higher insurance premiums for a specified benefit package than will healthy people. This is called the "equivalence principle". Individual health insurance may not be affordable for those at high risk if premium differences among individuals are rather extreme. It is, therefore, a major challenge to combine efficiency and financial access to coverage in the health system reforms that take place in many countries.

In practice, risk-rated premiums have been observed to differ across subgroups of insured people which are defined by rating factors such as age, gender, family size, geographic area, occupation, length of contract period, the level of deductible, health status at time of enrollment, health habits (smoking, drinking, exercising) and — via differentiated bonuses for multi-year no-claim — to prior costs (Van de Ven et al. 2000). Financial transfers are needed in order to avoid problems of financial access to coverage for those at high risk. The first and best solution in this case is to find a so-called sponsor who organizes compensation for those at high risk by setting up a regulatory system of risk-adjusted premium subsidies (Van de Ven et al. 2000). In the aforementioned European countries, the role of the sponsor is played by a government agency that organizes a system of risk-adjusted premium subsidies in the form of risk equalization among health insurers.[1]

A sponsor may not want to subsidize all premium rate variation observed in practice. For example, if premiums are rated across geographic areas, a sponsor may desire to equalize observed premium rate variation only up to the extent that this variation is caused by regional health status differences and not by regional differences in input prices. In general, the total set of risk factors that insurers use to rate their premiums can be divided into two categories: the subset of risk factors that causes premium rate variation which the sponsor decides to subsidize, the S(ubsidy)-type risk factors, and the subset that causes premium rate variation which the sponsor does not want to subsidize, the N(on-subsidy)-type risk factors (Van de Ven and Ellis 2000, p. 768-769). In most countries, up to a certain

1. See the glossary for a broader definition of the term sponsor.

extent, gender, health status, and age will probably be considered as S-type risk factors. Examples of potential N-type risk factors are a high propensity for medical consumption, living in a region with high prices and/or overcapacity resulting in supply-induced demand, or using providers with an inefficient practice-style (Van de Ven et al. 2000). The sponsor determines the specific categorization of S-type and N-type risk factors. Ultimately, if the sponsor is a national government, this categorization is determined by value judgments in society.

The risk-adjusted premium subsidies should compensate for cost variation caused by the S-type risk factors alone. Given the specific categorization of S-type and N-type risk factors, adequate measures of these S-type risk factors should be found in order to be able to implement a system of risk-adjusted premium subsidies in practice. It often turns out to be quite a challenge to find such adequate measures of the S-type risk factors at the individual level for the total population of insured people. Although age and gender of insured people may be easily implemented, the empirical possibilities to find adequate measures of health status are often limited because of feasibility concerns and a complex political and legal environment in which the scheme must operate. Amongst others, feasibility implies that the health status measure be available routinely for all insured people, both healthy and unhealthy.

In the absence of adequate measures of the S-type risk factors, it may be the case that, in practice, incomplete and/or imperfect measures of the S-type risk factors are used instead to calculate the risk-adjusted premium subsidies. For example, working status may be used as a measure of the S-type risk factor health status, although cost differences between employees and self-employed people may be partly caused by an N-type risk factor such as timeprice ("no time to visit a doctor") and the resulting propensity for visiting a doctor. An open question is, then, "To what extent do these incomplete and/or imperfect measures of the S-type risk factors induce risk-adjusted premium subsidies as intended by the sponsor?" The question of whether the premium subsidies are effective or not often proves difficult to answer in practice due to a lack of relevant data.

Under the approach developed in this study, an alternative system of risk-adjusted premium subsidies is defined for a limited sample of insured people. Because these risk-adjusted premium subsidies are calculated for a limited sample, an extensive set of measures of the S-type risk factors can be collected from additional data sources. For example, a tailor-made health survey can be conducted under a limited sample of insured people such that the health status profile can be described more precisely than with the limited range of health status measures used to calculate the actual risk-adjusted premium subsidies. The reason for this

is that the restrictions of feasibility are less stringent in this case. Therefore, this alternative system of risk-adjusted premium subsidies can be set up such that it better reflects the policy goals of the sponsor than the actual system of risk-adjusted premium subsidies.

For the limited sample of insured people, the actual system of risk-adjusted premium subsidies can be compared to the alternative system which is based on this broad array of measures of the S-type risk factors. This alternative system is normative in nature because it reflects the norms of the sponsor as accurately as possible. This normative system of risk-adjusted premium subsidies functions as a benchmark against which the workings of the actual system of risk-adjusted premium subsidies and that of alternative model specifications can be assessed. Notice that the approach developed in this study crucially depends on an adequate definition of this normative system of risk-adjusted premium subsidies.

A crucial difference with previous studies on systems of risk-adjusted premium subsidies is that, traditionally, the focus has been on incentives for selection with respect to subgroups of insured people, which may result from premium rate restrictions (Van de Ven et al. 2000) or transaction costs (Newhouse 1984). Under this traditional approach, the predictable profits and losses are calculated for these subgroups without paying attention to the question of whether these are caused by S-type or N-type risk factors (or a combination thereof). This type of research is especially relevant in situations in which insurers are not allowed to or not capable of risk-rating their premiums, which is the case in all of the aforementioned countries (for example, under community rating). However, in this study, it is assumed that insurers are fully free and capable of adjusting their premiums to an individual's risk. The competitive health insurance market is assumed to be regulated only in the sense that there is a periodic open enrollment requirement for a standardized benefit package and a system of risk equalization among health insurers.

The research strategy proposed in this study is relevant for all sponsors who need to determine whether a system of risk-adjusted premium subsidies functions in accordance with their policy goals. It may be applied in practice on a regular basis in all countries where a system of risk-adjusted premium subsidies is implemented in a competitive health insurance market.

Against this background, the purpose of this study is to develop a procedure for testing the effectiveness of the risk-adjusted premium subsidies to Dutch enrollees in 2004. Furthermore, alternative specifications of the risk equalization model and premium rate restrictions are studied in order to determine whether (and

how) these subsidies can be improved. In other words, the central question of this study is:

"To what extent does the 2004 Dutch risk equalization model induce risk-adjusted premium subsidies that meet the stated policy goals of the Dutch government and (how) can these subsidies be improved?"

In the next section, the Dutch regulatory system of risk-adjusted premium subsidies is explained in more detail. The policy goal of the Dutch government is to compensate insurers for a level of costs that is predictably higher than average as far as this is caused by age, gender and other objective measures of health status of the insured population (MoHWS 2005, p. 23).[2] Throughout this study, it is therefore assumed that the Dutch sponsor considers age, gender and health status to be the only S-type risk factors.

1.2 Risk equalization in the Netherlands

In 2006, the Dutch government effectuated a convergence between social health insurance, private health insurance, and the statutory scheme for civil servants and the police forces into a mandatory standard insurance policy for all 16 million inhabitants.[3] This convergence implies that since then, there exists one universal system of private health insurance under social conditions for non-catastrophic risks for all Dutch citizens.[4] This reform towards regulated competition in the Dutch health insurance sector is along the lines of Enthoven (1978).

Private health insurance means that people are entitled to the benefits as specified in the individual health insurance contracts with the private insurance companies they are enrolled with. Individual health insurance is mandatory under

2. Another policy goal of the Dutch government is to align financial risk of health insurers with their possibilities to control health care costs by application of an ex-post compensation scheme (MoHWS 2005, p. 26). The effects of a specific form of ex-post risk sharing will be presented in Section 7.2.

3. See MoHWS (2004) for an overview of the Dutch health insurance arrangements before 2006, and Lamers, Van Vliet and Van de Ven (2003) for a detailed description of the characteristics of the 2001 Dutch sickness fund sector. In 2004, about 64 percent of the Dutch population is enrolled at one of the 22 competing insurers for mandatory social health insurance. See Table A1.1 for a complete list of the Dutch sickness funds (2004).

4. Costs of expensive or long-term health care are covered for all Dutch residents under the Exceptional Medical Expenses Act (AWBZ), both before and after 2006. These costs are outside the scope of this study.

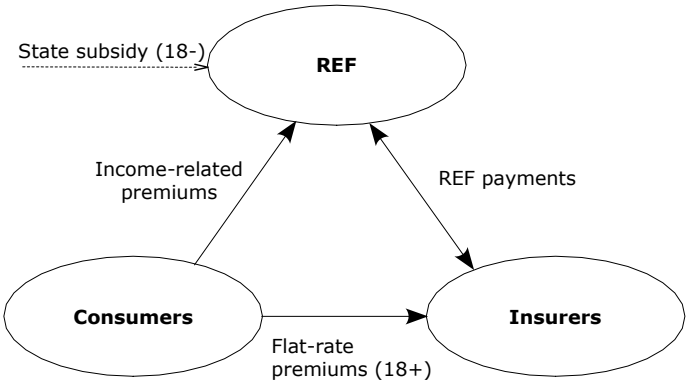


Figure 1.1: Financing system for the Dutch health insurance sector

the Dutch 2006 Health Insurance Act. Competition is regulated in the sense that a Risk Equalization Fund ("REF") is set up to distribute the risk-related cross-subsidies between the sick and healthy people. Additional regulatory measures taken by the government are that premium differentiation is only allowed between the 12 Dutch provinces, and is forbidden with respect to all other possible risk rating factors. Furthermore, there is an annual period of open enrollment during which enrollees are allowed to switch freely between insurers which are obliged to offer full insurance, but enrollees may opt for a voluntary deductible of 500 euro at most.[5]

Figure 1.1 shows that the premium subsidies are financed through the income-related contributions that all Dutch health insurance beneficiaries have to pay. The income-related contributions are levied by the tax collector and are received by the REF that redistributes them as risk-adjusted premium subsidies among the Dutch health insurers.

The payments that are channeled to insurers are equal to predicted costs minus a fixed per capita amount that is the same for all Dutch insured and is determined by the Dutch government. For enrollees younger than 18, this fixed per capita amount equals zero by law. This fixed withhld may be interpreted as the premium that the Dutch government expects an average insurer to charge in order to make ends meet. Of course, an insurer may charge a higher or even lower flat-rate premium than is indicated by the fixed withhold. Such deviations may arise be-

5. Paolucci, Den Exter and Van de Ven (2006) claim that premium rate restrictions and open enrollment should be avoided because they reduce efficiency and are unnecessary, not proportional, and undesirable to the pursuit of the general good. Premium and excess-loss compensation are second-best alternatives to the first and best alternative of REF models to guarantee an "acceptable level of solidarity"

cause the insurer is a better than average prudent purchaser of health care, for example.[6]

Enrollees younger than 18 are financed by a state subsidy to the REF. By Dutch law, half of the health care expenditures should be covered by income-related premiums and half by flat-rate premiums on average. As a consequence, there are subgroups of enrollees for which the payments to the insurer turn out to be negative. Therefore, in Figure 1.1 the arrow also depicts the possibility of a reverse payment flow from the insurers to the REF.

The REF model that holds under the new law is founded on the model that was set up and refined in the social health insurance sector before 2006. By law, about two-thirds of all Dutch citizens were enrolled in a sickness fund of their choice each year.[7] Risk equalization has been an essential element of gradual market-oriented health care reforms in the Dutch sickness fund sector since 1993. The risk adjusters that are chosen by the REF for risk equalization purposes are called REF adjusters. As the set of the REF adjusters was severely limited for reasons of administrative feasibility, only demographic variables such as age and gender were used in the early 1990s. Since 1995, eligibility status and regional area of residence were included as well but more direct health measures were not added until 2002.[8]

Although these demographic models compensate for predictable differences in health care expenditures among enrollees to some extent, they are not adequate enough as structural health status differences still remain among individuals who receive the same premium subsidy. In the absence of better health-based risk adjusters, ex-post reimbursements to reduce the financial consequences played a major role in the Dutch REF model during these years.

As of 2002, the Dutch REF model for sickness funds has been extended with more direct health measures, called outpatient pharmacy-based costs groups (PCGs), which identify enrollees suffering from specific conditions on the basis of

---

6. Note that absolute differences between the flat-rate premiums of different insurers are independent of the fixed amount subtracted by the REF. Relative differences are smaller, however, the larger the subtracted fixed amount.

7. Social health insurance was mandatory under the Health Insurance Act (ZFW) for those inhabitants being (self)employed and earning an income below a fixed gross income threshold. Their family members were also included. The same held for those receiving social benefits. People of 65 years of age with an income falling below a threshold could choose for this type of insurance on a voluntarily basis. See Table A1.1 for an overview of the 22 Dutch sickness funds in 2004.

8. See Table A1.2 for a detailed overview of the Dutch REF formulas for sickness funds over the years 1991 up to 2001.

outpatient prescribed drugs. Since 2004, the Dutch REF model also includes so-called inpatient diagnostic costs groups (DCGs), which identify enrollees suffering from specific conditions on the basis of inpatient diagnoses determined during hospitalizations.[9]

Currently, the Netherlands is the only country in the world that includes both the PCGs and DCGs in the REF formula and according to our knowledge it is therefore the most sophisticated formula in use at the moment. However, in addition to the prospective premium subsidies, the Dutch insurers are still reimbursed retrospectively for some part of their expenditures. In fact, 90% of the annual health care expenses above a threshold of € 12,500 are retrospectively reimbursed from the REF (i.e. a combination of "outlier risk sharing" and "proportional risk sharing").[10] The reason for this is that the current REF adjusters are still perceived to be too crude to reflect the health status differences to the full extent.

The downside of retrospective reimbursements is that they reduce efficiency. Therefore, the Dutch government wants to improve the ex-ante REF model, in order to be able to increase insurers' risk to strengthen the incentives for efficiency. For example, increasing the outlier risk sharing threshold of € 12.500 would increase financial risk beyond the 52% that holds in 2004. However, it is still an open empirical question how much of the necessary risk-related cross-subsidies between the healthy and sick people have already been achieved with the current set of REF adjusters, and which amount of individual variation in health status is still unexplained and should be addressed before insurers can be put at full risk.

The analysis in this study is based on 2004 data from a Dutch sickness fund. Application of the methodology proposed in this study may prove especially valuable in the context of the convergence in Dutch health insurance that took place in 2006. The reason for this is that the restriction of administrative feasibility that now limits the set of REF adjusters appears to be tighter than before 2006. In addition, the health status differences within the total Dutch population are larger than within the subpopulation of the sickness fund members. Therefore, more effort is needed in order to safeguard access to coverage for the high risk enrollees under the new Health Insurance Act.

---

9. See Table A1.3 for a detailed overview of the Dutch REF formulas for sickness funds over the years 2002 up to 2005.

10. The so-called "fixed" (i.e. production-independent) hospital expenses are excluded from this retrospective reimbursement arrangement.

1.3 Research questions

The first research question reads:

1. Given the definition of the basic benefits package, how can we calculate the risk-adjusted premium subsidies such that they meet the policy goals of the Dutch government? (Chapters Three, Four and Five)

In this study, it is assumed that the sponsor desires risk-adjusted premium subsidies for health care costs that are caused by the S-type risk factors age, gender and health status. We will refer to these costs as the "normative costs". Guided by the literature, administrative and health survey variables are made available for this study which reflect these risk factors best. In Chapters Three and Four, the data are described that are necessary to apply the test procedure proposed in this study; in Chapter Five, the normative costs are derived given the available data.

The second research question reads:

2. To what extent can the risk-adjusted premium subsidies be aligned with the policy goals of the Dutch government by means of the risk adjusters included in the 2004 Dutch REF equation? (Chapter Six)

The normative costs are compared with REF predicted costs from the 2004 Dutch REF model. In this way, it is determined to what extent these measures (for example, insurance eligibility and region) induce risk-adjusted premium subsidies as intended by the sponsor. In order to remove a potential gap between REF predicted costs and normative costs, a procedure is developed to adjust the regression weights of these REF adjusters such that desired risk-adjusted premium subsidies are better captured. The outcomes of this procedure are contrasted with those of the omitted variables approach for adjusting regression weights as proposed by Schokkaert, Dhaene and Van de Voorde (1998) and Schokkaert and Van de Voorde (2000, 2004). In chapter two, this alternative estimation procedure is described in more detail. Note that at the start of the present study, the 2004 model was the most recent year for which the REF model was known. The main difference between the 2004 model and the 2007 model is that the latter is applied to all 16 million Dutch citizens which made a redefinition of the REF adjuster "insurance eligibility" necessary.

The third research question reads:

3. To what extent can the risk-adjusted premium subsidies be aligned with the policy goals of the Dutch government by alternative specifications of the Dutch REF model or by premium rate regulation? (Chapters Seven and Eight)

From a claims dataset – derived from the administration of the Dutch health insurer Agis – substantially more administrative variables are available than are currently included as risk adjusters in the Dutch REF model. Predicted costs after adding this new set of administrative variables are compared to normative costs in order to determine the extent to which the risk-adjusted premium subsidies induced by the current Dutch REF adjusters can be improved upon.

From the above exercise, it should be clear to what extent the extended set of risk adjusters induce risk-adjusted premium subsidies that meet the policy goals of the Dutch government. However, it may still be possible that this goal is not achieved to the full extent. As a supplement to imperfect REF adjusters, ex-post risk sharing between health insurers and the Risk Equalization Fund may ultimately close the gap with normative costs in that case. A risk sharing scheme analogous to that applied in the 2004 Dutch REF model will be tested.

As a final illustration of the test procedure proposed in this study, the linear specification of the Duch REF formula is tested against an alternative specification that assumes (1) a multiplicative relationship between predicted health expenses and the risk adjusters and (2) heteroscedasticity in the error terms. The assumption of multiplicativity was dropped from the REF model with the introduction of PCGs in 2002. The assumption was that the PCGs capture cost differences more directly than a multiplicative specification of demographic risk adjusters alone. Furthermore, the stepwise and cell-based approach was replaced by the method of least-squares under the assumption of homoscedastic error terms. Both the multiplicativity and homoscedasticity assumptions are tested via the test procedure developed in this thesis.

1.4 Outline

In Chapter Two, the challenge to combine efficiency and financial access to coverage in competitive health insurance markets is described in more detail. Risk-related cross-subsidies among insurers to equalize cost differences are usually seen as the best strategy to make coverage affordable for those at high risk. In this study, it is assumed that the sponsor desires risk-adjusted premium subsidies for observed cost differences that are caused by the S-type risk factors age, gender, and health status. Guided by the literature, the administrative claims

data and health survey variables are chosen that reflect these risk factors best. The selection of risk adjusters to include in the normative risk equalization model answers the theoretical part of the first research question.

In Chapter Three, the data sources used in the calculations for this study are described. Firstly, the claims data are described which are derived from the administration of the Dutch health insurer Agis and sampled from 1997 up to 2002. Secondly, the construction of the Agis Health Survey 2001 containing the Dutch version of the SF-36 questions is described in terms of the administration mode, included survey questions, the data collection process and a response-nonresponse analysis. Thirdly, data from other, external sources (e.g. the Dutch consultancy companies APE and Prismant) are described.

In Chapter Four, the Agis Health Survey 2001 will be validated, mainly based on the 2001 administrative claims data. Although the Dutch SF-36 questionnaire has already been validated in Aaronson (1998) to some extent, our richer database makes it possible to validate at a more detailed level. Furthermore, the SF-36 health scales will be constructed and validated. The eight health status subscales, together with the physical and mental component summary scales will be derived. In order to capture health status differences as precisely as possible, the eight underlying health status subscales will be employed in the normative risk equalization model instead of the summary component scales. Furthermore, these subscales have the advantage of being more actionable, i.e. the scale scores are directly related to real-life situations such that concrete actions follow from the results in order to achieve better health outcomes.

In Chapter Five, an empirical answer to the first research question will be given. First, the administrative and health survey variables chosen in Chapter Two are described in a statistical way and cross-tabulated in order to check whether the motivation for inclusion also holds for the current study sample. After this validity check, the normative costs are derived as being the costs predicted by the chosen set of administrative and health survey variables.

In Chapter Six, the extent to which the set of risk adjusters used in the 2004 Dutch REF equation generate risk-adjusted premium subsidies that meet the policy goals of the Dutch government is determined. First, normative costs are compared to REF predicted costs given the 2004 Dutch REF model. In order to close the gap between the REF predicted costs and normative costs, an adjustment of the regression weights associated with the REF adjusters is tested. Alternatively, the omitted variables approach as proposed by Schokkaert, Dhaene and Van de Voorde (1998) and Schokkaert and Van de Voorde (2000, 2004) is applied to adjust the REF weights. This gives an answer to the second research question.

In Chapter Seven, a new set of risk adjusters is added to the risk adjusters already included in the 2004 Dutch REF model in order to determine to what extent they generate risk-adjusted premium subsidies that better meet the policy goals of the Dutch government than the original model. Secondly, an analogue of the 2004 Dutch risk sharing scheme is tested, which essentially boils down to a 90% retrospective reimbursement of actual health care costs above a threshold of € 12,500. Thirdly, predicted costs are derived in the context of a GLM framework under the assumption of a Gamma error distribution and a log link between predicted costs and the REF adjusters.

In Chapter Eight, it is determined to what extent premium rate regulation creates implicit cross-subsidies across subgroups of insured people for cost variation caused by S-type risk factors on the one hand and for cost variation caused by N-type risk factors on the other hand. This exercise is performed for subgroups of insured people defined by insurance eligibility, self-reported prior utilization, self-reported health status, diseases and conditions, claims-based prior costs and the regional categorization of Dutch provinces. These results and those of the three applications in Chapter Seven give an answer to the third research question.

In Chapter Nine, the conclusions are drawn and a discussion follows on potential future applications of the test procedure proposed in this study.

Appendix A1.1: Company names of Dutch sickness funds

Table A1.1: The 22 Dutch sickness funds (2004)

Company name	Legal name	CVZ id
Agis	OWM Agis Zorgverzekeringen UA	7
Amicon	OWM Amicon Zorgverzekeraar Ziekenfonds UA	127
AnderZorg	Onderlinge Ziekenfonds Maatschappij AnderZorg UA	43
Azivo	OWM AZIVO Algemeen Ziekenfonds De Volharding UA	54
AZvZ	Stichting Algemeen Ziekenfonds voor Zeelieden	13
CZ Groep	Stichting Centrale Zorgverzekeraars Groep, Ziekenfonds	119
De Friesland	OWM De Friesland Zorgverzekeraar UA	84
Delta Lloyd en OHRA Ziekenfonds	OWM Delta Lloyd en OHRA Zorgverzekering UA	53
DSW	OWM Zorgverzekeraar DSW UA	29
Geové zorgverzekeraar	OWM Geové zorgverzekeraar UA	65
Groene Land PWZ	OWM Groene Land PWZ Zorgverzekeraar UA	91
Nederzorg	OWM Ziekenfonds Nederzorg UA	45
ONVZ	Onderlinge ONVZ Ziekenfonds UA	38
OZ	OWM OZ zorgverzekeringen UA	22
OZB	OWM Onafhankelijk Ziekenfonds Bedrijven UA	44
Salland	OWM Salland zorgverzekeringen UA	32
Stad Rotterdam	OWM SR-Zorgverzekeraar UA	37
Trias	OWM Zorgverzekeraar Trias ua	50
Univé	Onderlinge Verzekerings Maatschappij Univé Zorgverzekeraar UA	1
VGZ	Stichting Ziekenfonds VGZ	95
Zilveren Kruis Achmea	OWM Zilveren Kruis Ziekenfonds UA	100
Zorg en Zekerheid	OWM Zorgverzekeraar Zorg en Zekerheid ua	85

APPENDIX A1.2: THE DUTCH REF MODELS FOR SICKNESS FUNDS 1991 – 2005

Table A1.2: The Dutch REF models for sickness funds 1991 – 2001 (Dutch guilders [a])

Year	Compensation variables / functional form	Outlier Risk Sharing (Dutch guilders)[a]	Pro-portional risk sharing	Retrospective reimbursement				Financial risk	Flat-rate premium pipy, adults only (Dutch guilders) [a]		
				Out-patient	Prod. Dep. Hosp.	Med. Spec.	Prod. Indep. Hosp.		MoHWS premium withhold	Mean	Band-width
1991	historical costs	- - -	- - -	100% [b]	100% [b]	100% [b]	100% [b]	0%	220	226	0
1992	+ age/gender	- - -	95%	100% [b]	100% [b]	100% [b]	100% [b]	0%	171[c]	198	6
1993	age/gender	- - -	90%	75%	75%	75%	75%	3%	171	198	6
1994	idem	- - -	90%	75%	75%	75%	75%	3%	171	198	6
1995	+ region [d] + disability	- - -	90%	75%	75%	75%	75%	3%	171	198	6
1996	+ redefinition of hospital costs [e]	- - -	60%	50%	50%	50/95% [f]	95%	15%	267 [c]	344	36
1997	+ disability*age [g]	90% ≥ 4500 [h]	30%	25%	25%	25/95%	95%	27%	157 [i]	216	95
1998	+ specialist care separately [f]	idem	30%	15%	25%	95%	95%	29%	157 [i]	216	95
1999	+ insurance eligibility*age [j]	90% ≥ 7500	30%	0%	25%	95%	95%	35%	296	398	96
2000	+ historical costs [k] + ORS-correction	90% ≥ 10000	0% [k]	0%	25%	95%	95%	36/50% [l]	310	425	147
2001	idem [m]	90% ≥ 10000 [n]	0/50% [o]	0%	25%	40%	95%	38/53% [l]	324	362 [p]	201

[a] The 2002 euro conversion rate equals 2.20371 Dutch guilders.

[b] Ex post agreed upon between MoHWS and the ZN branche organisation of Dutch health insurers.

[c] As of 1992, January 1st, prescribed medicines are no longer part of the sickness funds benefit package. As a consequence, lower flat-rate premiums are charged since then. As of 1996, January 1st, prescribed medicines are part of the sickness funds benefit package again (categorized as outpatient costs) and flat-rate premiums increased as a consequence.

[d] Region is classified according to degree of urbanization of municipalities (as of 1999, January 1st, based on four-digit ZIP codes).
[e] Total hospital costs are split up in a variable part (i.e.: production-dependent) to which the risk measures are applied, and a fixed part (i.e.: the production-independent hospital costs) that is divided between sickness funds based on their historical costs and 95% of the financial result is reimbursed retrospectively (no proportional risk sharing). In 1996, 60% proportional risk sharing and 50% retrospective reimbursement is applied to the total of variable hospital costs and outpatient costs.
[f] Since 1998, costs of specialist care (private practitioners and those on the payroll of a hospital) are a separate cost category of which 95% of the financial result is reimbursed retrospectively (no proportional risk sharing). Retrospective reimbursement of the financial result with respect to outpatient costs is dropped to 15%. In 1996 and 1997 private practitioners of specialist care are de facto categorized as generating outpatient costs, whereas costs generated by specialists on the payroll of a hospital are partly treated as variable hospital costs and partly as fixed hospital costs.
[g] Subgroups of disabled workers and non-disabled workers are both differentiated into five age categories.
[h] ORS (Outlier Risk Sharing) is for the sum of costs of outpatient care and production-dependent hospital care (in 1997 including medical specialist care).
[i] In 1997 and 1998, a coinsurance rate of 10% with a yearly maximum of 200 Dutch guilders was applicable for most types of care offered by sickness funds. This is the reason why the flat-rate premium is relatively low in those years.
[j] "Insurance eligibility" has the following categories (globally): disabled worker, employed, on welfare, unemployed and pensioner.
[k] In 2000, proportional risk sharing is replaced by a historical cost component of 30%, based on the sickness fund specific average of costs per enrollee over the years 1996-1998, separately for outpatient costs and variable hospital costs, and corrected for changes in the sickness fund enrollees' composition according to age/gender, insurance eligibility/gender and region.
[l] If the historical component is treated as a kind of "ex post risk sharing" then the financial risk in 2000 equals 36%; however, if the historical component is treated as a risk measure on its own, then the financial risk equals 50%. With respect to 2001, these figures amount to 38% and 53%, respectively.
[m] See Appendix A in Lamers, Van Vliet and Van de Ven (2003) for a detailed description of the Dutch system of risk adjustment and risk sharing in the year 2001.
[n] The division of the costs to be pooled between variable hospital costs and outpatient costs is determined at the level of the individual enrollee. Before 2001, it was calculated at the level of the sickness fund.
[o] Besides 40% retrospective reimbursement, also 50% proportional risk sharing holds with respect to specialist care in 2001 (proportional risk sharing is calculated per enrollee, before 2001 per premium-equivalent).
[p] The reduction of the average flat-rate premium is related to the cap on financial reserves of sickness funds set by the government since 2001.

Table A1.3: The Dutch REF models for sickness funds 2002 – 2005 (Euro [a])

Year	Compensation variables / functional form [b]	Outlier risk sharing (euro) [a,c]	Outpatient	Prod. Dep. Hosp. + Med. Spec.		Prod. Indep. Hosp.	Financial risk	Flat-rate premium pipy, adults only (euro) [a,d]		
			Historical costs	Pro-portional risk sharing	Retro-spective reim-bursement			MoHWS premium withhold	Mean	Band-width
2002	age / gender, insurance eligibility [e] / age, APE-region [f], PCGs [g]	90% > 7500	30%	30%	35%	95%	41 / 52% [h]	155	183	125
2003	idem, self-employed and employed in same risk category [i]	idem	0	30%	35%	95%	52%	257	348	151
2004	self-employed as separate risk category + DCGs	90% > 12500	0	30%	35%	95%	53%	222	305	143
2005	idem	idem	0	30% [j]	35% [j]	95% [k]	53%	72 [l]	383 [m]	216

[a] The 2002 euro conversion rate equals 2.20371 Dutch guilders.

[b] Since 2002 the REF formula is entirely linear in its administrative REF adjusters, i.e. additive instead of multiplicative. Before 2002 insurance eligibility / age and region were included as multiplicative risk measures with respect to the base REF predicted costs, which depended on age / gender). Furthermore, the methodology to split total hospital costs into a fixed part and a variable part is changed since 2002.

[c] ORS (Outlier Risk Sharing) is applied to the sum of the costs of outpatient care, production-dependent hospital care and medical specialist care.

[d] Since 2001, April 1st, the flat-rate premium is only charged for adult enrollees (i.e. 18 years and older). Before 2001, for children half of this flat-rate premium had to be paid.

[e] "Insurance eligibility" has the following categories (globally): disabled worker, employed, on welfare, unemployed and pensioner.

[f] APE-regions: a classification of four-digit ZIP codes into 10 categories by the APE consultancy firm, The Hague, The Netherlands.

[g] PCG (Pharmacy Cost Groups): a classification of enrollees into disease groups derived from pharmacy use of specific drugs (for more than 180 calendar days) in the year before.

[h] If the historical component is treated as a kind of "ex post risk sharing", the financial risk in 2002 equals 41%; however, if the historical component is treated as a risk measure on its own, the financial risk equals 52%.
[i] Since 2000 self-employed enrollees with a "small" firm size are obliged to get sickness fund insurance instead of private health insurance. From 2000 till 2003, self-employed enrollees fall into the same eligibility category as employed enrollees. Since 2004, self-employed are a separate eligibility category.
[j] Financial results with respect to the fixed hospital costs are fully reimbursed instead of 95%, if it concerns so-called "academic component" costs.
[k] Since 2005, hospital claims are based on so-called Diagnosis and Treatment Combinations (DTCs), a system of hospital output prices that is no longer based on administrative budgets and financing rules that is the Dutch alternative to the Diagnosis Related Group (DRG) concept. Due to a lack of historical data, it is impossible to take account of the financial consequences of this change in hospital claims structure in the REF model for 2005. To reduce the financial risks for sickness funds, a 90% retrospective reimbursement of sickness funds' financial results is applied (after outlier risk sharing, proportional risk sharing and regular retrospective reimbursement) if outside a bandwidth of 1.5% around a specified reference result. This reference result is a combination of the percentage financial result in 2005 averaged over all sickness funds, and the sickness fund specific historical financial result as far as it concerns the variable hospital care and specialist care cost categories.
[l] This relatively low MoHWS premium withhold can be explained by a mandatory markup of the premium by 255 euro for all Dutch insurers because of a so-called no-claim arrangement. Since 2005, a no-claim arrangement is introduced to raise adult patients' awareness about the costs of health care and to reduce the consumption of care (cost containment argument). Those adult enrollees who do not consume care will receive their no-claim back at the end of the year, up to a maximum of 255 euro. For each visit to a health care provider and in case of prescribed drugs and medical devices, the no-claim restitution of 255 euro is reduced by some amount. GP visits and visits to an obstetrician or maternity nurse are excluded from this no-claim arrangement. On average, MoHWS (2004) expects a 93 euro no-claim restitution by the end of year 2005.
[m] This is an estimation by MoHWS (2004), including the mandatory 255 euro no-claim markup and excluding the 93 euro expected no-claim restitution (see footnote l).

Chapter 2

THEORETICAL FRAMEWORK AND METHODS

2.1 A TEST FOR THE EFFECTIVENESS OF RISK-ADJUSTED PREMIUM SUBSIDIES

2.1.1 Risk-adjusted premium subsidies in competitive health insurance markets

During the past few decades, health system reforms have taken place in many countries in order to achieve an optimal distribution of resources in terms of price and quality of health care services. In the absence of any additional regulation and as long as there are no entry barriers, premiums *per individual contract* tend to equal expected expenditures in such competitive health insurance markets, i.e. the so-called "equivalence principle". In our framework, it is assumed that insurers are fully free and capable of rating their premiums according to risks.[11]

Many rating factors may be used to adjust premiums for systematic variation in actual expenditures across individuals.[12] Figure 2.1 summarizes the broad range of potential rating factors into seven general classes of health insurance risk factors: age and sex, health status, socio-economic characteristics, provider characteristics, input prices, market power of the insurer and benefit plan characteristics.

The first three classes of risk factors depicted in Figure 2.1 are characteristics of individuals (Van de Ven and Ellis 2000): age and sex, health status, and socio-economic factors, such as lifestyle, taste, purchasing power, religion, race, ethnicity, and population density. The fourth group includes all provider characteristics, such as practice styles and whether there is an oversupply of providers or facilities. Input prices are a characteristic of the region in which the providers are located and are largely exogenous to the patient and provider. The final two groups are characteristics of the insurer. Market power depends on the insurer's ability to negotiate price discounts. Benefit plan features include conventional demand side features, such as deductibles, co-payments and decisions about covered services, but also include supply side features, such as utilization reviews, various health management strategies, characteristics of contracts and financial

11. In this study, in contrast to Newhouse (1996), it is assumed that insurers have sufficient information at their disposal to accurately adjust the premiums to a consumer's risk; this also includes high-risk consumers and new applicants. Therefore, transaction costs do not hinder insurers from differentiating their premiums.

12. For example, in the Netherlands, Van de Ven et al. (2000) observed an increasing differentiation of premiums for individual private health insurance plans during the last quarter of the previous century. By the end of this period, rating factors being applied included age, gender, family size, region, occupation, length of contract period, individual versus group contract, the level of deductible, health status at time of enrollment, health habits, smoking, drinking, exercising and — via differentiated bonuses for multiple years of no claims — prior costs.

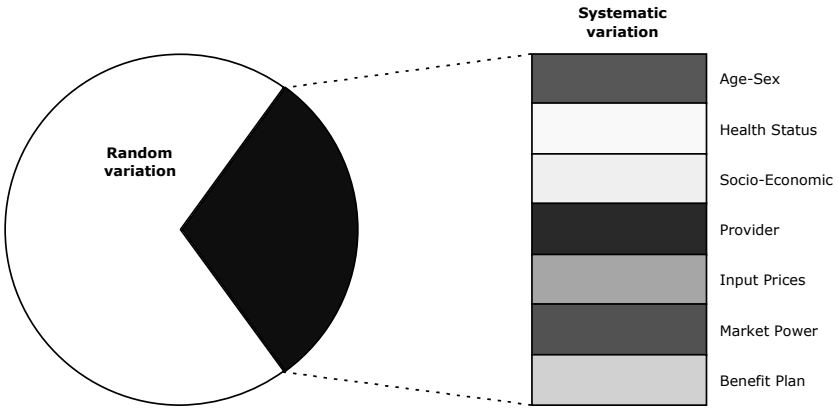


Figure 2.1: Seven classes of risk factors explaining variation in health care costs (Van de Ven and Ellis 2000)

incentives between plans and providers.[13]

As a consequence of the equivalence principle that holds for competitive health insurance markets, healthy people pay lower insurance premiums than sick people. Premium differences across individuals may be rather extreme. The maximum premium for full coverage (i.e. without cost-sharing) could be expected to exceed the minimum premium for the same product by more than a factor of 100 (Van de Ven et al. 2000). In such a market, efficiency incentives for competing insurers are at their maximum, but premiums may become unaffordable for sick people.

In this study, it is assumed that the sponsor wants to safeguard affordability for high-risk individuals. This can be done through cross-subsidies from low-risk to high-risk individuals that are organized by a sponsor – for example, the government – and distributed via a so-called Risk Equalization Fund (REF). The subsidies are in the form of risk-adjusted premium subsidies. These risk-based subsidies are preferred over cost-based subsidies that reduce incentives for efficiency and are also preferred over premium-based subsidies that entail reduced incentives for high-risk consumers to shop around for the lowest premium, an increase of moral hazard resulting from over-insurance and a misallocation of subsidies to the extent that premium rate differences observed among insurers reflect cost differences caused by N-type risk factors (Van de Ven 2006).[14]

13. Note in Figure 2.1 that substantial random variation in health care costs remains ex-ante, even after controlling for these seven classes of risk factors. This remaining expenditure variation will be averaged out by insurers through risk pooling.

14. Van de Ven (2006) rejects the organization of means-tested, tax-financed subsidies as proposed by Zweifel and Breuer (2006). The advantage of non-means tested, risk-adjusted premium subsidies (possibly funded by means-tested taxes) is that large unpredictable fluctuations in out-of-pocket

Theoretically, these risk-adjusted premium subsidies should be based on so-called acceptable costs. Acceptable costs are defined as the costs of services that follow from a quality, intensity and (demand and supply) price level of treatment that the sponsor considers to be acceptable to be subsidized. For example, if the sponsor defines the level of costs generated in delivering only medically necessary and cost-effective care to be acceptable, then the cost of hospitalization will not be taken into account when calculating the risk-adjusted premium subsidies if only day surgery is medically indicated. Note that the specification of the standardized benefit package is supposed to be given in this definition of acceptable costs.

In practice, 'acceptable costs' are hard to determine, because this requires a time-consuming case-by-case study of the acceptable level of treatment for all health care consumers. The premium subsidies are therefore usually based on health care costs actually observed in practice, for example, in the Netherlands.[15] However, although risk-rated premiums tend to capture *all* systematic cost variation observed in competitive health insurance markets, in this study, it is assumed that the norm of the sponsor is that risk-adjusted premium subsidies should only compensate for cost variation among subgroups of insured people caused by so-called S(ubsidy)-type risk factors (Van de Ven and Ellis 2000, p. 768-769).[16] Normative costs are then defined as the observed, actual costs that follow from this sponsor's norm. So-called N(on subsidy)-type risk factors capture cost variation for which no cross-subsidization is desired.

The selection of S-type risk factors plays a crucial role in the scientific and political debate. In all countries with risk equalization schemes, there is a consensus that ideally, subgroups included in the risk equalization formula should be related to the health risks of the insured population. In other words, the sponsors in these societies desire cross-subsidies between their healthy and sick people. According to Van de Ven and Ellis (2000), most societies desire cross-subsidization for age, sex and health status. Under the 2006 Dutch Health Insurance Act, the purpose of the REF model is to compensate for differences in health status among insured people that are caused by age, sex, and objective measures of health status

annual premiums (i.e. premium minus subsidy) can be avoided and transaction costs of means-testing for both consumers and government can be reduced.

15. The same approach holds for social health insurance programs, such as Medicare in the USA or sickness fund systems in Germany and Israel (Van de Ven and Ellis 2000).

16. Of course, other norms are possible, too. For example, a sponsor may require the system of risk-adjusted premium subsidies to be based upon the pattern of cost variation that is observed among a preferred group of health care providers under the assumption that these providers deliver only medically necessary and cost-effective care.

(MoHWS 2005, p. 23).[17] In this study, it is therefore assumed that the Dutch government desires risk-adjusted premium subsidies for the cost differences among subgroups of enrollees defined by the S-type risk factors age, sex and health status.[18] In general, it may be difficult to calculate risk-adjusted premium subsidies as intended by the sponsor due to reasons described in the next section.

2.1.2 Causes for incomplete and imperfect REF adjusters

Premium subsidies compensate adequately if the subgroups defined by the S-type risk factors are accurately included in the REF equation. However, in practice, the subgroups are defined by proxies of S-type risk factors, also called REF adjusters. Although the S-type risk factors age and gender may be easily measured in practice, this often proves difficult with the S-type risk factor health status. Potentially, there are a lot of direct and indirect proxies of health status that one can think of, but if these variables are included in the REF equation, they will not always satisfy the criteria of effectiveness, appropriateness, and feasibility that should hold for REF adjusters in general:

- Effectiveness of risk-adjusted premium subsidies: The REF adjusters induce risk-adjusted premium subsidies that adequately compensate insured people for cost differences caused by S-type risk factors, and do not compensate for cost differences caused by N-type risk factors;
- Appropriateness of incentives: Risk-adjusted premium subsidies should not reduce the incentives for efficiency or health-improving activities; there should be no incentives for distorting information used to calculate risk-adjusted premium subsidies;[19]
- Data feasibility: REF adjusters should be routinely available at reasonable costs and the system should be acceptable to all parties involved (for example, there should be no conflict with privacy).

The general criteria of effectiveness, appropriateness, and feasibility guide the choice of REF adjusters (Van de Ven and Ellis 2000). The criterion of effective

17. The need for a society to make its goals explicit is in accordance with the WHO recommendations to make societal goals for countries explicit (Murray and Frenk 2000).

18. Input prices are also likely to be considered an S-type risk factor in some societies.

19. In contrast to Van de Ven and Ellis (2001), incentives for selection are not listed here. Within the context of this study, insurers are allowed to risk-rate their premiums, and therefore, incentives for regulation-induced selection (and their adverse effects) are not relevant. Under the assumption that transaction costs do not hinder insurers from substantially differentiating their premiums, transaction cost-induced incentives (and their adverse effects) for selection are also absent (cf. Newhouse 1996).

risk-adjusted premium subsidies covers the main purpose of any REF model: to induce a level and direction of risk-adjusted premium subsidies among insured people that compensate for differences in normative costs. This criterion is only fully satisfied if the subgroups defined by the REF adjusters and those defined by the S-type risk factors are identical.

Ideally, risk-adjusted premium subsidies should not reduce incentives for efficiency in the production of health care. However, these incentives are reduced if (some proxy of) prior costs or prior utilization is included as a risk adjuster in the REF equation. The reason for this is that inefficiencies from the past might be rewarded via risk-adjusted premium subsidies. So, to the extent that REF adjusters reduce incentives for efficiency, there exists a tradeoff between this reduction and the effectiveness of the risk-adjusted premium subsidies.

In practice, the choice of REF adjusters is usually severely restricted because of feasibility problems. If any adequate measures of health status at the individual level can be made available, it may turn out that these are only available for a limited part of the insured population. Subgroups that are defined by S-type risk factors, but which can not be made available for inclusion in the REF equation, are insufficiently compensated. The consequence of such an incomplete set of REF adjusters is that S-type cost variation will not be fully captured in practice. Therefore, risk-adjusted premium subsidies will not adequately compensate high-risk individuals.

The aforementioned criteria of appropriateness of incentives and data feasibility limit the choice of REF adjusters, in which case the criterion of the effectiveness of the risk-adjusted premium subsidies may not be fully satisfied. This implies that risk-adjusted premium subsidies do not adequately compensate insured people for cost differences between the subgroups defined by the S-type risk factors. Another type of problem may be that cost variation among subgroups defined by these REF adjusters may also be caused by N-type risk factors. In econometric terms, the REF weights are biased and inconsistent. This bias will persist even for large samples. The REF adjusters are called imperfect in this case.

Figure 2.2 illustrates the bias that occurs if a REF adjuster compensates for cost variation caused by S-type risk factors as well as N-type risk factors. The vertical axis represents observed costs and the horizontal axis depicts the S-type risk factor health status in the case of the "True relationship" and the REF adjuster health status in the case of the "Observed relationship". Suppose that the "True relationship" between health care expenditures and good health is negative, which is reflected by the negative slope of the population regression line drawn in this figure. This population regression line defines the level of normative costs

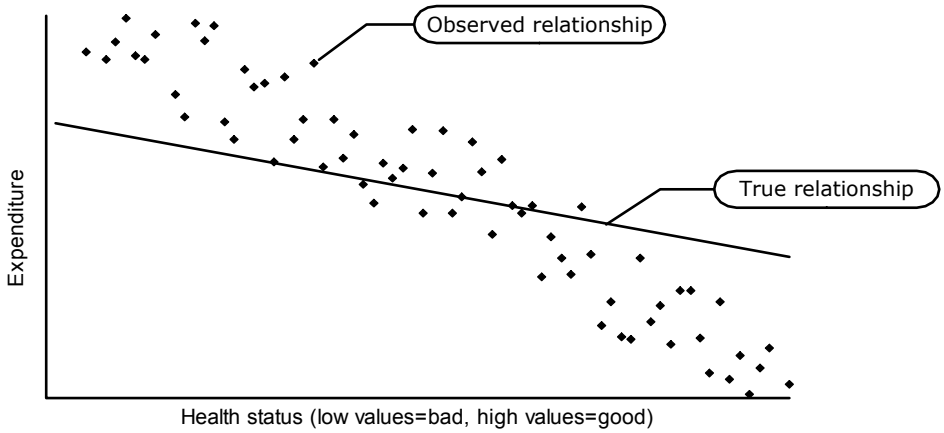


Figure 2.2: Biased REF weights as a result of imperfect risk adjustment

for various health status levels, but this relationship is usually not observed in practice.

The population regression line would be estimated correctly if the REF adjuster would have been a perfect measure of the S-type risk factor health status. The crucial point here is that all the researcher ever observes is the scatter of points that reflects an empirical relationship between the imperfect REF adjuster health status and observed costs ("Observed relationship"). As a consequence, the REF adjuster may wrongfully capture N-type cost variation at the same time that it captures S-type cost variation. For example, the health status effect will be overestimated if a REF adjuster not only captures health status but also regional supply effects to some extent, while supply and good health are negatively correlated. In terms of Figure 2.2, this means that the estimated empirical slope is larger (in absolute value) than the true slope of the population regression line.

Imperfect REF adjusters, due to problems with limited availability, are very common in practice. Take, for example, eligibility for insurance membership, which is included as a risk adjuster in the 2004 Dutch REF equation. One of the subgroups defined by eligibility consists of those members who are self-employed. Before 2004, there was much debate on whether or not to include self-employment as a REF adjuster. Given that these members are associated with relatively low predicted costs, there are two competing explanations for this phenomenon: (1) self-employed members are in relatively better health due to some self-selection processes, and (2) given their health status, they are faced with relatively high opportunity costs (i.e. time-price) to get the care they need. If the first explanation holds, the subgroup of self-employed members should be included in the

REF equation to remove potential premium differences. If the second explanation holds, this subgroup should not be included.

In the end, the Dutch government argued that cost differences caused by self-employment most probably resulted from a better health status and did include this subgroup in the REF equation. However, it was acknowledged that self-employment probably reflects both S-type and N-type cost variation. In other words, given the health status, low health care utilization might result partly from a relatively better health status and partly because of relatively high opportunity costs. In this case, the decision to include this subgroup in the REF equation induces more risk-adjusted premium subsidies than desired by the sponsor. At the same time, it also reduces incentives for insurers to act as prudent purchasers of care for the self-employed and/or to charge a lower premium by using "self-employment" as a risk-rating factor.[20] It would be more appropriate in this case to compensate for expenditures only up to the extent that self-employment reflects health status differences. The conclusion is that with imperfect measures of health status, such as self-employment, there exists a problem in separating the specific part of the cost differences for which insurers have to be compensated from that for which insurers should be held responsible.

Another REF adjuster about which much discussion exists is regional variation in expenditures. Potential causes of the observed regional variation in expenditures are health status differences. Given the health status, other causes include differing practice styles, taste differences, excess supply, regional variation in costs of living or differences in access across regions. If health status differences are the cause of these regional variation in expenditures, then risk equalization might be seen as fair and the region must be included in the REF equation. No compensation should be offered for the other causes listed above, thus treating regional variation in expenditures as N-type cost variation. In the latter case, the insurer has an incentive for efficiency and/or may differentiate premiums across regions.

However, just as in the case of self-employment, the observed pattern of health care expenses predicted by the REF adjuster "region" may not be attributable to health status differences alone. It may not be the exclusive result of non-health status-related phenomena either. The conventional dichotomous procedure to include regional variation in expenditures as a REF adjuster is inappropriate in this case. Actually, the more relevant policy question is for which part there is S-type

20. The option to charge a lower premium using "self-employment" as a risk-rating factor is only viable under the assumption made in this study that insurers are fully free and capable of risk rating their premiums. However, in practice, risk rating premiums is legally forbidden in the Netherlands.

regional cost variation and for which part N-type variation exists. This question is answered in this study.

In the next subsection, methodological solutions are described to correct the regression weights of imperfect REF adjusters if they are biased by N-type risk factors. Given the set of REF adjusters, a correction of biased REF weights does not necessarily improve the actual risk-adjusted premium subsidies in capturing S-type cost variation. In the subsection thereafter, methodological solutions for improving cross-subsidization for S-type cost variation are discussed.

2.1.3 Methodological solutions for imperfect REF adjusters

The conventional approach to find an estimate of normative costs on which the premium subsidies can be based is to estimate a REF equation by ordinary least squares. The REF equation is of reduced form if the included REF adjusters are all exogenous measures of the S-type risk factors.

In case of imperfect REF adjusters, such as insurance eligibility and region, the estimated REF weights may also capture cost variation caused by N-type risk factors. To remove this bias from the REF weights, Schokkaert, Dhaene and Van de Voorde (1998) and Schokkaert and Van de Voorde (2000, 2004) advocate the procedure of including measures of cost variation exclusively caused by N-type risk factors during the estimation phase and exclude these measures when calculating risk-adjusted premium subsidies. More precisely, the effects of the N-type risk factors are set at some level desired by (or: 'acceptable' to) the sponsor, which is the same for all insured, usually at average population values.

This alternative estimation technique implies a change from using a reduced form modeling strategy to a structural form modeling strategy, in which case the bias in the REF weights can be interpreted as a so-called omitted variables bias. The purpose of this structural form approach is to separate the effects of S-type and N-type risk factors on expenditures during the estimation phase by an adjustment of the REF weights. In terms of Figure 2.2, the intention of this procedure is to find the slope of the true relationship between normative costs and the REF adjusters.

Ideally, applying an omitted variables bias approach would remove the bias from the REF weights corresponding to self-employment and the regional risk adjuster in the example of the Dutch REF equation above. However, the proposed solution for the omitted variables problem is not applied in practice in most countries because of problems of limited feasibility and methodological problems. The problem of limited feasibility exists because implementing a REF model along these lines puts an additional burden on market parties. In addition to REF adjusters that should

capture S-type cost variations, following this procedure implies that variables that capture N-type cost variation should be collected at the individual level as well.

Methodological problems arise because unbiased and consistent structural form parameter estimates are only obtained if the included measures of the N-type risk factors are exogenous, i.e. they are not correlated with the measured and the unmeasured parts of normative costs. However, N-type cost variation may be correlated with past and current values of the S-type risk factors, a so-called problem of endogeneity in econometric terms. This may be the case because, historically, funding has been based explicitly on the (regional) distribution of the insured according to the S-type risk factors in resource allocation formulas. N-type risk factors are then correlated with current values of the (observed and unobserved) S-type risk factors if the S-type risk factors are stable over time (see also Gravelle et al. 2003). This leads to biased coefficients in the REF equation if estimated by the ordinary least squares technique and therefore, to an improper estimation of the REF weights.[21],[22]

Schokkaert and Van de Voorde (2004) expect the endogeneity problem to be negligible at an individual level of analysis, provided that the N-type risk factors are measured at a higher level of aggregation. They reason that it is not realistic to assume that individual medical expenditures will influence medical supply at the regional level, for example.[23] Carr-Hill et al. (1994, 3.19) acknowledge that much of the endogeneity of the N-type risk factors depends on the level of analysis being adopted.

As a solution to a potential endogeneity problem, Carr-Hill et al. (1994) apply two-stage least squares methods to estimate the parameters of their structural form model for costs at a small area level of analysis. In order to successfully apply this approach, instruments for the supply variables should be obtained that do not correlate with normative costs, neither the measured nor the unmeasured

---

21. More generally, if both risk factors S* and N* are unobserved and N is a perfect measure of N*, then the bias in the coefficient corresponding to S* can be entirely removed by including N during the estimation phase. However, if N is an imperfect measure of N*, then OLS estimates are inconsistent. In this case, it might be better to use OLS without N measures (Wooldridge 2001, p. 64).

22. If the unmeasured part of the cost variation caused by S-type risk factors is correlated with the measured part as captured by the REF adjusters, the effect of the unmeasured part may be correctly picked up by the estimated REF weights. This may be a valid reason to exclude supply variables during the estimation phase (see also Schokkaert and Van de Voorde 2004).

23. However, this argument does not necessarily hold if regional heterogeneity in S-type risk factors is correlated with regional heterogeneity in N-type risk factors, even if equation (2.3) is estimated at the individual level.

part. After an optimal set of variables is collected and included in a full-fledged structural form model, a reduced form equation is estimated for the selected REF adjusters by simply ignoring the instruments for the N-type risk factors from the structural form equation. Estimated REF weights in this reduced form capture both direct and indirect effects of the S-type risk factors. However, finding such instruments often proves to be a difficult, if not impossible, task. For example, Gravelle et al. (2003) allow for the endogeneity of supply variables by including regional dummy variables at a higher level of aggregation. The construction of such a multi-equation structural form model, also proposed by Schokkaert and Van der Voorde (2004), is not pursued in this study.

Another methodological solution for removing the bias from the REF weights is possible if normative costs are made observable for a limited subsample of insured people according to the procedure developed in this study. The difference with the REF model that must apply to the total population of insured people is that with respect to this subsample, the limitations of feasibility are less severe. Therefore, a broad array of health status measures can be made available at the individual level for this subsample of insured people by means of a health survey. This broad array of variables constitutes a more precise measure of the S-type risk factor health status than in case of a limited set of REF adjusters. Therefore, a proper estimate of normative costs can be derived, albeit for this subsample of insured people only.

Deviations from normative costs can be attributed to N-type risk factors. Such bias in the REF weights can be avoided by using normative costs instead of observed costs as the dependent variable when estimating the REF equation. This allows adjusted REF weights to be used instead of the original REF weights when calculating the premium subsidies. In Section 2.3 it will be demonstrated that applying this adjustment procedure will completely remove the bias from the REF weights by construction. Notice that, although based on the limited subsample of survey respondents, these adjusted REF weights can also be applied to calculate premium subsidies for the total population of insured people in practice.

2.1.4 Methodological solutions for incomplete REF adjusters

Given the calculation of normative costs for a limited subsample of survey respondents, the extent to which the REF equation under study generates effective risk-adjusted premium subsidies is determined by a comparison of REF predicted costs with normative costs for the subgroups defined by the S-type risk factors. If equality holds for all of these subgroups of insured people, then the REF equation fully satisfies the criterion of effectiveness of the risk-adjusted premium subsidies.

However, under the assumption made in this study that a limited set of REF adjusters is used in the REF equation, this criterion will not be fully satisfied.

The specification of the REF equation should be improved upon in order for the risk-adjusted premium subsidies to become more effective. For example, new risk adjusters could be made available to add to the REF equation, ex-post risk sharing may be applied (Van Barneveld 2000, and Van de Ven et al. 2000), or the functional specification of the REF equation might be changed.[24] The procedure that is used to test the REF equation can also be used to test alternative specifications for the effectiveness of the risk-adjusted premium subsidies as well. Furthermore, if an alternative specification of the REF model is implemented in practice, any remaining bias caused by N-type risk factors can be removed from the REF weights according to the adjustment procedure described in the previous subsection.

Given some specification of the REF model, the criterion of effectiveness may still not be fully met for the subgroups of insured people defined by the S-type risk factors. To improve the risk-adjusted premium subsidies, often additional regulations, such as premium-rate restrictions and open enrollment, are implemented into practice. However, such regulations create incentives for selection that threaten the quality, affordability, and efficiency of care. This is described in more detail in the next subsection.

2.1.5 Additional regulations for improving the subsidies

As far as the risk-adjusted premium subsidies insufficiently compensate high-risk individuals, they can be improved by the additional regulatory measure of restricting out-of-pocket premium rates (i.e. premium minus subsidy). A combination of premium subsidies and such regulations is used in several countries (Van de Ven et al. 2000).[25] The least restrictive type of rate regulation is rate-banding, where premium rates are restricted to vary between a low-rate and high-rate band, as set with respect to an index rate by the sponsor. A special case of rate-banding occurs if the sponsor requires uniform pricing, i.e. the sponsor effectively sets the low-rate band equal to the high-rate band. Rate-banding and community rating

---

24. If the REF equation is supplemented by an ex-post risk sharing arrangement between insurers and the Risk Equalization Fund, it comes at the expense of insurers' incentives for efficiency in production. Therefore, in case of ex-post risk-sharing, there is a trade-off between effectiveness of the risk-adjusted premium subsidies and efficiency. Note that a risk-sharing arrangement contrasts with traditional reinsurance because it is mandatory and the price for an insurer is not (fully) related to the risk of its members for whom some risk is shared.

25. In general, restrictions on rating practices apply to a consumer's direct payments to insurers, i.e. either the premium, the premium minus the premium subsidy, or the premium minus the premium subsidy plus the solidarity (e.g. income-related) contribution.

may hold for all individuals within certain risk classes defined by, for example, the geographic location, industry, family size or smoking history. Alternatively, rate-banding and community rating may hold, irrespective of the risk classes used by insurers when setting their premiums. This type of rate-banding is more restrictive as it reduces an insurer's ability to price their insurance risks when setting premium rates. Community rating per insurer per product can be seen as a very restrictive form of banning the use of rating factors as it implies that for each product, an insurer must ask the same premium from each enrollee, completely independent of the individual's risk characteristics. A sponsor may also decide to ban only a limited set of rating factors instead of requiring community-rated premiums.

Although the sponsor may decide to regulate premium rates in order to avoid S-type cost variation from being priced by insurers, at the same time implicit subsidies may be induced for N-type cost variation between the subgroups defined by the corresponding rating factor. Such subsidies are undesired as they conflict with the policy goals of the sponsor. For example, to the extent that the costs for those who are self-employed are lower than those who are employed as a result of N-type risk factors, such as time-price ('no time to visit a doctor'), this cost variation should not be equalized. Consequently, insurers should be allowed to give a premium rebate to the self-employed. The procedure developed in this study makes it possible to determine to what extent premium rate restrictions affect the pricing of S-type cost variation across subgroups defined by the rating factors and to what extent they restrict the pricing of N-type cost variation. From the perspective of the policy goals of the sponsor, (implicit) cross-subsidies induced by rate regulations can be judged as more effective when these cross-subsidies better capture S-type cost variation.

However, although premium rate restrictions are intended to have a positive impact by inducing implicit cross-subsidies from low-risk to high-risk individuals, at the same time, they create predictable profits and losses at the individual level and therefore incentives for selection. Selection is defined as actions (not including risk rating) by consumers and insurers that exploit unpriced risk heterogeneity and break pooling arrangements (Newhouse 1996). In the literature, two types of selection are distinguished: adverse selection and cream skimming. These forms of selection are different from each other in terms of the type of selection actions that may actually be undertaken by consumers and insurers, as well as in their effects on efficiency and affordability.

Adverse selection actions by consumers may arise if consumers have an information surplus over the insurers, which may be the result of government regulations (i.e. premium rate restrictions) on the health insurance market or because

of asymmetric information between consumers and insurers in unregulated competitive health insurance markets (Wilson 1977). Adverse selection in regulated markets may either cause a competitive insurance market to be unstable (i.e. an ongoing entry and exit of market parties) or it may result in a pooling or separating equilibrium (Rothschild and Stiglitz 1976, Wilson 1977, Schut 1995, and Newhouse 1996). In the latter case, high-risk individuals pay a high premium for generous health insurance coverage and low-risk individuals pay a low premium for stingy coverage (Van de Ven and Ellis 2000).[26],[27]

Cream skimming is a type of selection selection that occurs because insurers prefer low-risk consumers to high-risk consumers within the same premium risk group (Van de Ven and Ellis 2000). Cream skimming (or "preferred risk selection", or "cherry picking") may be undertaken by insurers in regulated health insurance markets if they have an information surplus over the REF. There are financial incentives for cream skimming if the profits of these actions outweigh its costs. In unregulated health insurance markets, cream skimming may also exist because transaction costs of risk rating may be too high.[28]

Premium rate restrictions may lead to cream skimming activities that threaten quality, affordability and efficiency in care (Van de Ven, Van Vliet and Lamers 2004). First, insurers have a disincentive to respond to preferences of high-risk consumers. Consequently, the chronically ill might receive poor quality care or poor service. Although the findings in the literature do not warrant final conclusions, insurers could encounter strong financial incentives to be unresponsive to preferences of the chronically ill. Second, to the extent that some insurers attract low-risk consumers, these selection activities result in market segmentation, wherein high-risk patients are enrolled with insurers that charge high premiums and low-risk patients self-select into insurers with low premiums. That is, selection may threaten affordability. Third, at least in the short run, selection may be more profitable than improving efficiency and therefore efforts to improve efficiency may be suboptimal. In summary, restrictions on premium contributions that are intended to induce implicit cross-subsidies at the same time provide incentives for selection that may threaten quality, affordability, and efficiency in care.

---

26. A straightforward way of preventing an extreme form of adverse selection, i.e. low risk individuals who do not subsidize high risk individuals as they buy no coverage at all, is to mandate that everyone buys the specified health insurance coverage.

27. Note that asymmetric information may increase adverse selection (and moral hazard).

28. Newhouse (1996) claims that selection is only induced by transaction costs that preclude plans' pricing at an individual's expected costs. Van de Ven et al. (2000) argue that selection may also be induced by premium rate restrictions.

Van de Ven and Van Vliet (1992) note that, although denying coverage may be legally forbidden, subtle and hidden risk selection strategies may prove successful. An example of such strategies may be medical screening in supplementary health insurance policies that indirectly limits enrollees' ranges of choice with respect to their basic insurance policies, if basic and supplementary insurance policies are (perceived to be) sold as a one-package deal.

Regulating a health insurance market may also lead to volume rationing and quality skimping, even if perfect risk adjustment has removed all incentives for preferred risk selection. Van de Ven and Schut (1994) mention two types of care for which quality skimping may exist: care that is used by people who do not have the mental ability to make a tradeoff between price and quality themselves and care in which most people are not interested because they have a very low probability of needing it during the next contract period. In this study, it is assumed that these types of care are not included in the basic benefits package.

To conclude, the goal of the REF is to channel explicit cross-subsidies from low-risk to high-risk individuals, but the risk-adjusted premium subsidies may turn out to be incomplete. In practice, there often exist additional regulations, such as premium-rate restrictions, that generate implicit cross-subsidies from low-risk to high-risk individuals. This, however, creates incentives for selection that threaten quality, affordability, and efficiency in care. The better the cross-subsidies are adjusted for the S-type risk factors, the less additional regulation is needed.

The contribution of this study is the development and application of a procedure to test the effectiveness of risk-adjusted premium subsidies that result from REF models and to remove any bias caused by N-type risk factors from the REF weights. In section 2.2, the international literature on risk adjusters is reviewed. These risk adjusters are candidates for inclusion in the normative equation for the survey respondents. In section 2.3, the methodological issues of this new approach to risk equalization will be treated more in-depth and in a technical manner. Furthermore, the equations to be estimated will be formulated and a guideline is presented for the interpretation of the results when this test procedure is applied in the empirical part of this study.

2.2 Risk adjusters in the literature

In theory, the best strategy for safeguarding financial access to coverage in competitive health insurance markets is to use adequate measures of the theoretical S-type risk factors in order to adjust the premium subsidies. In practice, however,

the risk adjusters included in the REF equation most often appear not to fully satisfy the criteria of effectiveness of the risk-adjusted premium subsidies, appropriateness of incentives, and data feasibility (Van de Ven and Ellis 2000). This often results in complicated tradeoffs among these criteria (Van de Ven 2001). In this section, an overview is given of the risk adjusters that are currently in use in REF equations or under study for future implementation.

Demographic variables

The most basic type of information used in REF equations around the world are age and sex. Age and sex satisfy the three criteria of effectiveness of the risk-adjusted premium subsidies, appropriateness of incentives, and feasibility, but according to Van de Ven and Ellis (2000), their most serious drawback is simply that they are relatively weak predictors of individual expenditures. Note that from the normative point of view taken in this study, age and sex are adequately measured risk factors for which risk-adjusted premium subsidies are desired.

Prior year expenditures

Prior year costs may be added to demographic measures in REF equations in order to improve cross-subsidization between healthy and sick people. The inclusion of prior year costs in the REF equation leads to a substantial improvement over demographic-only equations (Van Vliet and Van de Ven 1992, and Ash et al. 1998). Predictive power is comparable to that achieved by diagnosis-based equations or equations that use self-reported health status measures.[29]

Although prior year expenditures may improve cross-subsidization between sick and healthy people, they may also lead to higher compensation as a result of prior care provided in an inappropriate way (McClure 1984) or provided by poorly managed insurance companies (e.g. Lubitz 1987, and Porrell and Turner 1990). Furthermore, one-off acute conditions may lead to inadequately high compensation during the next year (Beebe et al. 1985). Finally, using prior utilization/expenditures as a risk adjuster is not effective from the perspective of individuals that suffer from medical problems but who have not yet sought care. Note that these three arguments against inclusion in the REF equation do not hold with respect to demographic risk adjusters, such as age and sex.

Van de Ven and Ellis (2000) state that the first argument against inclusion of prior costs as a risk adjuster in the REF equation misses the point that insurers

29. Van de Ven and Ellis (2000) mention a feasibility problem in the USA, as a growing number of health plans do not collect individual level cost information that can be used for calculating payments for specific conditions. In the Netherlands, however, this feasibility requirement is met.

are still only compensated for a portion of their spending on health services. Newhouse et al. (1989), Van de Ven and Van Vliet (1992) and Ash et al. (1998) find a portion of about 20-30%, i.e. spending an extra dollar on health care in year one predicts extra spending of $0.20 and $0.30 in year two. The inclusion of prior year costs therefore reduces the incentive to only contain costs. Ellis and McGuire (1993) and Newhouse (1996, 1998) have argued that this may be a desirable practice to soften the incentives for selecting a fully prospective system. Marchand, Sato, and Schokkaert (2003) study the trade-off between efficiency versus selection in the context of a proportional ex-post risk sharing scheme, and demonstrate that it always improves welfare when prior costs are included as a risk adjuster. The crux here is that with ex-post risk sharing, observed costs are partially reimbursed anyway, which is certainly (although perhaps marginally) worse from the perspective of efficiency.

Diagnosis-based risk adjustment

The potential of inappropriate incentives associated with prior utilization can be reduced if combined with diagnostic information. The three most widely known classification systems are:

- The Ambulatory Care Group (ACG) system, developed at Johns Hopkins by Jonathan Weiner and colleagues [Weiner et al. (1991, 1996)];
- The Diagnostic Cost Group (DCG) family of models, developed at Boston University and Health Economics Research by Arlene Ash, Randall Ellis, Gregory Pope and colleagues [Ash et al. (1989, 1998); Ellis et al. (1996a, 1996b); Pope et al. (1998a, 1998b, 1999)];
- The Disability Payment System (DPS), developed by Richard Kronick, and Anthony Dreyfus, [Kronick, et al. (1996)] primarily for U.S. Medicaid disabled enrollees.

All diagnosis-based REF equations are based on the premise that clinical homegeneous diseases entail homogeneity in treatment and therefore, in associated costs. As such, they may be valid predictors of health status-induced cost differences. Since there are approximately 15,000 valid International Classification of Diseases (ICD-9) codes, each of the models listed above therefore begins by grouping ICD-9 codes into more aggregated groups, based on clinical, cost, and incentive considerations. The approaches that each model uses, and the way that information is used to generate predictions, differ in the three models listed above (Van de Ven and Ellis 2000).

Since 2000, Medicare premium subsidies in the USA made by the Centers for Medicare & Medicaid Services (CMS) are health-based, including so-called Principal In-Patient Diagnostic Cost Groups (PIP/DCGs) in the Medicare risk equalization

formula, and are explicit indicators of health status differences among enrollees (Ash et al. 1989, Ellis and Ash 1995, and Pope et al. 2000). Before 2000, Medicare premium subsidies to private health care plans were set at 95% of average adjusted per capita cost (AAPCC), calculated for enrollees of traditional FFS schemes.

PIP-DCGs are based on the "worst" diagnosis recorded as the principal reason for hospital admission during a one-year base-period, i.e. the diagnosis "having the highest future cost implications". Since 2004, CMS has moved to the so-called Hierarchical Condition Category (HCC) DCG model, which recognizes the cumulative effect of multiple health problems (Ellis et al. 1996, Ash et al. 2000, Pope et al. 1998, and Pope et al. 2004). This DCG/HCC model is an "all encounter" model because data are used from several sites of service, whereas the PIP-DCG model is based on data from inpatient diagnoses alone.

Since 2004, principal inpatient DCGs are also included in the Dutch REF equation. Where Van Vliet and Van de Ven (1993) applied some preliminary versions of the PIP-DCG model, Lamers (1998) applied the Ash et al. (1989) version that forms the basis for the Dutch REF equation. Lamers (1999b) assesses the predictive accuracy of the DCG equation by using survey information. Lamers and Van Vliet (1996) suggest including multiyear diagnostic information as another possibility for improving the risk equalization scheme.

Information derived from prescription drugs

The potential of inappropriate incentives associated with prior utilization can also be reduced if combined with the use of outpatient prescription drugs. In the Netherlands, since 2002, so-called Pharmacy-based Cost Groups (PCGs) are included in the REF equation. Lamers (1999a) describes a preliminary PCG classification that is based on a revised version of the Chronic Disease Score by Clark et al. (1995), originally developed by Von Korff, Wagner and Saunders (1992). It is concluded that the use of information on chronic conditions derived from claims for prescribed drugs is a promising option for improving the system of risk-adjusted premium subsidies.

However, Lamers (1999a) and Ellis (1985) note the potential problem of inappropriate incentives, as the additional subsidy for a PCG-classified enrollee may exceed the costs of the prescribed drugs that form the basis for a PCG-assignment. Lamers and Van Vliet (2003) describe the best strategies for reducing these gaming possibilities:

- Use the number of prescribed daily doses to assign people to chronic conditions and set a high threshold in order to be eligible for PCG classification
- Assign people to one condition, and no more, in order to reduce incentives for prescribing additional drugs to someone already classified in a PCG

- Exclude chronic conditions that lead to relatively small increases in premium subsidies, such that the number of drugs for which prescriptions may be profitable is reduced.

In the Dutch REF equation, all three strategies against gaming have been applied from the start of the introduction of PCGs in 2002: enrollees are only eligible for classification if the number of defined (instead of prescribed) daily doses is such that prescribed drugs are meant to be taken for at least half a year; a hierarchical clustering procedure is applied in order to assign enrollees to the one PCG that generates the largest increase in the premium subsidies; only 14 out of the 23 original PCGs are included in the risk equalization formula. The 2002 version of the Dutch PCG classification is described by Lamers and Van Vliet (2004).

Self-reported health information

As an alternative to using information essentially based on judgments of medical professionals, health status measures can be sampled from the patient's perspective. These health status measures are typically collected through surveys and can be categorized as follows:

- Perceived health status
- Functional health status
- Chronic conditions
- Other self-reported measures

Table 2.1 provides examples of these health status measures.

In the USA, the Centers for Medicare & Medicaid Services (CMS) include the SF-36 as the main instrument for collecting health outcome data from Medicare beneficiaries in the Health Outcomes Survey (HOS). The HOS was initiated in 1996, from a recognized need to monitor the performance of managed care plans. The HOS data are also a major constituent of the Health Plan Employer Data and Information Set (HEDIS®). For a recent evaluation of the HOS program, see Jones, Jones, and Miller (2004). For an introduction to the HEDIS program, see NCQA (2005).

Thomas and Lichtenstein (1986) and Hornbrook and Goodman (1996) show that disability and functional health status are good predictors of future health care costs. Gruenberg et al. (1989) show that the impairment level is a significant contributor to high Medicare expenditures, even after controlling for demographic factors and prior utilization. Mobility impairments are the leading reason for functional limitations among adults (Iezzoni et al. 2001). Newhouse (1986) considers disability to be an almost ideal risk adjuster. In 2000, an indicator of disability is used as a risk adjuster in Belgium, Germany, and the Netherlands (Van de Ven et al. 2003).

Table 2.1: Classification of self-reported health status measures

Type of information	Examples
Perceived health status	A single self-reported health summary of excellent/very good/good/fair/poor
	Asking how health status has changed since the previous year
	More elaborate surveys, such as the Short Form 36 (SF36) [Thomas and Lichtenstein (1986); Ware and Sherbourne (1992)] or the closely related RAND-36 survey [Hornbrook and Goodman (1995)][a]
Functional health status	Activities of Daily Living (ADLs)
	Instrumental Activities of Daily Living (IADLs)
Chronic conditions	Diabetes
	High blood pressure
	Asthma
Other self-reported measures	Lifestyle (smoking, drinking, food)
	Marital status, employment
	Education

Source: Van de Ven and Ellis (2000)

[a] See Iezzoni (2003, p. 61) and Ware (1995, p. 330) for an overview of widely used general health surveys and the concepts they measure.

Given the observation that diagnosis-based risk adjusters do not fully predict the expenditures of those who are frail and elderly, Kautter and Pope (2005) demonstrate that adjusting the CMS-hierarchical condition categories (HCC) model for frailty by measuring functional impairments has added value. The measure consists of four classes of counts of difficulties in performing Activities of Daily Living (ADL). ADL is seen as the most promising functional status measure for frailty adjustment, however, in their case, ADL scores are only applied at the health plan level as they are not available at the individual enrollee level.

Self-reported health status measures have advantages as well as disadvantages when compared to the use of DCGs and PCGs in REF equations. These are presented in Table 2.2. In this study, the health status is measured by self-reported perceived health status (the SF-36 scales), functional health status (OECD) and the prevalence of a number of preselected chronic conditions. All listed advantages are relevant in the context of the approach developed in this study, with the exception of uniformity across health plans because, in this case, data are sampled for only one insurer. On the other hand, most disadvantages listed in this table appear not to be relevant in the context of the approach developed in this study. In Chapter Three, it will be shown that, compared to other data sources currently available, response rates with respect to the 2001 Agis Health Survey are quite acceptable

Table 2.2: Advantages and disadvantages of using self-reported health status measures, compared to using diagnostic and/or drug prescription information

Advantages	Disadvantages
Most information is not contingent on having come in contact with a medical provider	Surveys are relatively costly to collect
No prior history of claims or enrollment is needed to generate predictions	Response rates can be unacceptably low
Measurement of consumer perceptions of need and anticipated use	Response rates can be correlated with medical risk
Uniformity across health plans	Large samples on which to develop reliable prediction models generally do not exist
Measurement of socioeconomic (lifestyle, taste, employment) variables	Confidentiality and accuracy concerns (e.g. questions about HIV/AIDS or mental illness)
	Reliability and validity of data collection procedures (e.g. non-random sampling)
	Mostly lower explanatory power versus diagnosis-based systems

Source: Van de Ven and Ellis (2000)

and the number of records available for analysis is relatively large. Furthermore, the data collection procedure may be called reliable as it is designed according to scientific guidelines with respect to conducting mail surveys and validating the self-reported data. The survey data are non-randomly sampled, but the strata are weighted back to population proportions. Confidentiality is guaranteed, as both the data collection and data analysis procedures are explicitly designed according to Dutch privacy standards. The only disadvantage listed in Table 2.2 that also holds for this study is that of data collection costs, which are indeed non-negligible.

Note that the self-reported information used in this study is only applied to a limited subsample of Dutch insured. The advantages and disadvantages discussed above are more relevant if such information is sampled for all Dutch people who are insured for which the REF equation holds in practice.

Mortality

Van Vliet and Lamers (1998) conclude that mortality at the individual member level should not be used as a risk adjuster, a.o. because of its relatively low explanatory power in terms of R^2 at the individual level and its small effect on the allocation of resources at the level of sickness funds. Furthermore, receiving a higher subsidy in cases of a higher mortality rate might seem to entail perverse incentives. Beck and Zweifel (1998) advocate mitigating this incentive problem by determining compensation for the cost of death prospectively and reimbursing retrospectively. In Belgium, a risk adjuster for mortality recorded at the sickness fund level is included in the risk equalization formula (Schokkaert and Van de Voorde 2003).

Other variables

Other risk factors that might influence utilization are sociodemographic or socio-economic risk factors, behavioral risk factors and physiological risk factors. Based on predictive accuracy performance, Lamers (1997) only identifies physiologic measures as a good candidate for inclusion in a REF model. Another advantage is that physiologic functioning is strongly associated with the prevalence of chronic diseases. However, a major disadvantage of using such measures as REF adjusters is the additional administrative burden and costs associated with data collection and the periodic assessment of risk factors (Schauffler et al. 1992).

Iezzoni (2003) also mentions the extent and severity of co-existing diagnoses (i.e. comorbidities) as potential REF adjusters. In our study, comorbidities are taken into account as non-rank-ordered PCGs and DCGs, which are included in the normative equation that applies to the subsample of survey respondents. Furthermore, the total number of self-reported diseases is included in this equation. It is also possible to take comorbidities among all Dutch insured into account in the national REF model by using a non-rank-ordered version of the PCGs and DCGs instead of a rank-ordered version, but the added value of this exercise is not determined in the context of this study.

2.3 Methods

2.3.1 REF predicted costs as an approximation to normative costs

In empirical studies, a major challenge is to find adequate measures of the S-type risk factors to include in the REF equation. The measures that are actually included in the REF equation as a proxy of the theoretical S-type risk factors are called REF adjusters. Age, sex, and DCGs are amongst the most prominent of such REF adjusters in case the sponsor chooses for age, sex and health status as S-type risk factors. The weights of these REF adjusters may be estimated either by a cell-based approach or a linear regression approach.

A cell-based approach to estimate average costs per cell is preferred from a feasibility point of view, but only if the number of REF adjusters and the implied subgroups is limited. If the number of REF adjusters is large, the multi-dimensionality of the cross-table becomes problematic to handle in practice and estimations are therefore obtained by linear regression techniques instead. Such linear regression techniques make it possible to choose to ignore certain (or: all) interaction effects between the REF adjusters. The corresponding mathematical specification of the REF equation is given by:

$$(2.1) \quad Y_t = \alpha_{0,t} + \sum_{j=1}^{J} \alpha_{j,t} * X_{j,t-1} + \epsilon_t$$

where Y_t are the health care costs observed in year t, $X_{j,t-1}$ is the j^{th} REF adjuster observed in year t-1 which is an incomplete and/or imperfect measure of the S-type risk factors, j=1,...,J, $\alpha_{j,t}$ are unknown coefficients to be estimated and ε_t is an independent and identically distributed error term in year t. The selection of the $X_{j,t-1}$ is guided by the usual criteria of effectiveness of the risk-adjusted premium subsidies, appropriateness of incentives, and feasibility (see Section 2.1.2). The variables Y_t, $X_{j,t-1}$, and ε_t are Nx1 vectors, the elements of which contain the observations with respect to enrollees i=1,...,N. The index t with respect to the $\alpha_{j,t}$ coefficients indicates that these weights may not be constant over the years.

After the estimation of REF equation (2.1) by linear regression, REF predicted costs $\hat{Y}_t$ constitute an estimate of normative costs and can be calculated as follows:

$$(2.2) \quad \hat{Y}_t = \hat{\alpha}_{0,t} + \sum_{j=1}^{J} \hat{\alpha}_{j,t} * X_{j,t-1}$$

If a REF adjuster $X_{j,t-1}$ takes on discrete values only, which is often the case with risk equalization equations, then the set of N individuals used to estimate equation (2.2) can be partitioned into subgroups such that each subgroup contains individuals with the same value for $X_{j,t-1}$. As a consequence of a property of the ordinary least-squares technique, the following identity then holds for each such subgroup of insured people defined by the j^{th} REF adjuster $X_{j,t-1}$:

$$\frac{1}{n_j(x)} \sum_{i \in I_j(x)} \hat{y}_{it} \equiv \frac{1}{n_j(x)} \sum_{i \in I_j(x)} y_{it}$$

where $\hat{y}_{it}$ and y_{it} are the elements of the vectors $\hat{Y}_t$ and Y_t corresponding to individual i, respectively, $I_j(x)$ constitutes the set of indices i of the individuals for all of whom the REF adjuster $X_{j,t-1}$ takes on the value x, and $n_j(x)$ equals the number of individuals belonging to a subgroup defined in this way.[30] For each REF adjuster $X_{j,t-1}$, the sum of $I_j(x)$ over all possible values x is equal to the total set of indices {1,2,...,N}. In other words, average REF predicted costs for some subgroup defined by the REF adjusters are equal to average observed costs for this subgroup of insured people by construction. This identity does not necessarily hold for subgroups which are not defined by the REF adjusters. Furthermore, together with the zero-sum property of the ordinary least-squares technique, this implies

30. Note that the sum of $n_j(x)$ over all possible values x equals the total number of individuals N.

that cost variation caused by N-type risk factors for one of these subgroups will always lead to biased REF weights of at least one of the other subgroups. If the REF adjuster $X_{j,t-1}$ takes on non-discrete values x instead of discrete values, then the above mentioned property of the ordinary least-squares technique only holds for the total sample of N individuals, that is:

$$\frac{1}{N}\sum_{i=1}^{N}\hat{y}_{it} \equiv \frac{1}{N}\sum_{i=1}^{N} y_{it}$$

If the REF adjusters are incomplete and/or imperfect measures of the S-type risk factors, then REF predicted costs constitute an incomplete and/or imperfect estimate of normative costs. In the next subsection, it is demonstrated how normative costs can be determined more precisely for a limited sample of insured people in this case. Given normative costs, the effectiveness of the risk equalization model can be expressed in terms of how closely REF predicted costs follows normative costs.

2.3.2 The determination of normative costs

In the literature, the proposed way to quantify health care need within a general population is to monitor health status by health surveys. For the purpose of this study, a tailor made health survey is conducted in order to construct a broad array of more precise measures of the S-type risk factor health status for a subsample of insured people.

In order to guide the selection of health status measures to be used to derive normative costs, the conceptual model of Ruwaard and Kramers (1997) is used. This conceptual model is employed in all Dutch "Public Health Exploration of the Future" publications (e.g. Van Oers 2002). Four indicators of health status can be distinguished in this conceptual model: (1) diseases and disorders, (2) functioning and quality of life, (3) mortality, and (4) (un)healthiness and life expectancy. The

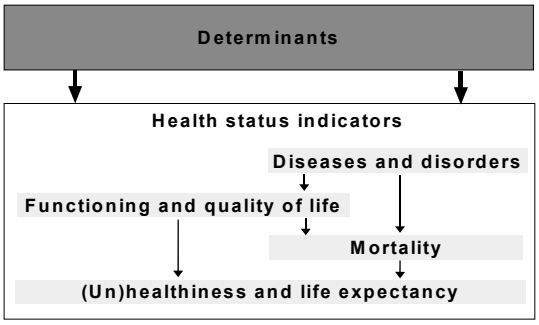


Figure 2.3: Four indicators of health status according to Ruwaard and Kramers (1997).

arrows in Figure 2.3 define the relation among these indicators. Diseases and disorders may have an influence on functioning and health status and on mortality, possibly indirectly via functioning and health status effects. (Un)healthiness and life expectancy is an indicator that can be seen as a combination of the three aforementioned indicators into one summary measure, thereby applying value judgments about the importance of the differences in outcomes of the summary measure between population groups.

In this study, health status will be defined in terms of three health indicators that determine "(un)healthiness and life expectancy" as shown in Figure 2.3, i.e. functioning and quality of life, specific diseases and disorders, and mortality. No use will be made of the indicator "(un)healthiness and life expectancy" itself, as for predictive purposes summarized valuations of health status are not appropriate in the proposed normative approach to risk equalization. Only measures of health status without any valuation thereof are included.

In this study, diseases and disorders are self-reported as well as derived from medical claims data (a.o. pharmaceutical prescriptions and ICD-9-CM hospitalization codes). Functioning and quality of life is derived from the answers to the SF-36 and the OECD questions, and mortality is defined as the expectation at the age, sex and municipality level.

The normative costs then result from a regression analysis where the expenditures are explained by explicit measures of the S-type risk factors age, sex, and individual health status, the so-called S-type adjusters. The reduced form equation can be described as follows:

$$(2.3) \quad Y_t = \beta_{0,t} + \sum_{l=1}^{L} \beta_{l,t} * S_{l,t-1} + \eta_t$$

where Y_t = costs observed in year t, $S_{l,t-1}$ = is the l^{th} S-type adjuster as observed in year t-1, l=1,..,L, $\beta_{l,t}$ are unknown coefficients to be estimated, and η_t is an independent and identically distributed error term in year t.

The PCGs and DCGs are also included in equation (2.3) for the following reason. In general one might expect a negative relationship between the SF-36 health status scores and the level of expenditures, i.e. medical care for the healthy will be less than for the unhealthy enrollees. However, to the degree that medical care influences health status and medical care is greater with more severe problems, this relationship may not necessarily be observed in the data. In fact, the SF-36 scores for people under treatment for a chronic disease may be similar to those for people without any disease, although the level of health care expenditures differs among these groups. Think e.g. of diabetes. Furthermore, within the group

of these patients under treatment, some may need more intensive treatment than others in order to arrive at the same health status score.

The same departure from the negative relationship between health status and medical expenditures holds if health status is measured by medical outcomes instead of SF-36 scores. Newhouse (1989) gives the example of a hypertensive individual whose blood pressure is controlled at 90 mmHg but whose uncontrolled value is at 105 mmHg. During the period of observation, this patient could be expected to have higher medical expenditures than an otherwise identical hypertensive individual who is not under treatment. Furthermore, within the group of controlled enrollees, treatment costs may be larger for an individual with an uncontrolled diastolic blood pressure of 110 mmHg than 100 mmHg because the case is more severe.

The conclusion is that variance in expenditures remains unexplained if the (intensity of the) treatment effect is not accounted for. A (negative) monotone relationship between the SF-36 scores and health care costs would be more probable if these are a reflection of intrinsic health status, i.e. if scores are observed independent from the treatment effect. As it is impossible to find these intrinsic health scores in practice, specific diseases and disorders (see Figure 2.3) must be added to equation (2.3) instead. These diseases and disorders may act as predictors of treatment expenditures that are not captured by the observed SF-36 health scores. Variation in expenditures caused by severity differences within treated groups of enrollees might be captured by estimating SF-36 weights separately for each included disease and disorder (Hornbrook and Goodman 1996). This approach is not pursued in this study.

After estimation of equation (2.3), the normative costs can be calculated according to the following formula:

$$(2.4)\quad Y_t^{NORM} = \hat{\beta}_{0,t} + \sum_{l=1}^{L} \hat{\beta}_{l,t} * S_{l,t-1}$$

If an S-type adjuster $S_{l,t-1}$ takes on discrete values, then the set of N individuals used to estimate equation (2.4) can be partitioned into subgroups such that each subgroup contains individuals with the same value for $S_{l,t-1}$. As equation (2.3) is estimated by ordinary least squares, normative costs are equal to observed costs on average for each subgroup defined by the l^{th} S-type adjuster $S_{l,t-1}$:

$$\frac{1}{n_l(s)} \sum_{i \in I_l(s)} y_{it}^{NORM} \equiv \frac{1}{n_l(s)} \sum_{i \in I_l(s)} y_{it}$$

where y_{it}^{NORM} and y_{it} are the elements of the vectors Y_t^{NORM} and Y_t corresponding to individual *i*, respectively, $I_l(s)$ constitutes the set of indices *i* of the individuals for all of whom the S-type adjuster $S_{l,t-1}$ takes on the value *s*, and $n_l(s)$ equals the number of individuals belonging to a subgroup defined in this way. If the S-type adjuster also takes on non-discrete values, then this property only holds at the level of the total sample of *N* individuals.

Deviations from normative costs may be observed for other subgroups than those defined by the S-type adjusters, which should be interpreted as underutilization if average observed costs fall short of normative costs and as overutilization otherwise. In mathematical terms, for a subset *Z* of individuals *i*:

$$\begin{cases} \Delta_t^Z < 0\text{: } \textit{underutilization} \\ \Delta_t^Z > 0\text{: } \textit{overutilization} \end{cases}$$

where

$$(2.5) \quad \Delta_t^Z = \frac{1}{n_Z}\sum_{i\in I_z} y_{it} - \frac{1}{n_Z}\sum_{i\in I_z} y_{it}^{NORM}$$

where I_Z constitutes the set of indices *i* of the individuals belonging to the subset *Z*, and n_Z equals the corresponding number of individuals.

Furthermore, normative costs summed over all insured equals the sum of observed costs by the same argument. Therefore, overutilization for some subgroup implies underutilization for some other subgroup(s) defined by an S-type adjuster. Note that this relative property of normative costs would not hold if the parameters in equation (2.3) are set equal to some predetermined desired level instead of being estimated by ordinary least squares.

Remember that the ordinary least squares estimation technique is also applied to REF equation (2.1) in order to derive REF predicted costs from equation (2.2). Therefore, REF predicted costs are equal to observed costs on average for each of the subgroups defined by the REF adjusters. As a consequence, REF predicted costs are equal to normative costs if and only if these subgroups defined by the S-type adjusters can also be defined by the REF adjusters.[31]

In order to determine the extent to which the REF equation generates risk-adjusted premium subsidies as intended by the sponsor, REF predicted costs should be compared to normative costs for the subgroups defined by the S-type adjusters.

31. For example, average REF predicted costs are equal to average costs for subgroups of enrollees formed on the basis of the REF adjusters age and gender. This is because age and gender enter in equation (2.2) as well as equation (2.4).

If equality holds for all such subgroups, then it must be concluded that the REF model fully satisfies the criterion of effectiveness. However, under the assumption made in this study that the set of REF adjusters is limited because of restrictions of feasibility, a gap between REF predicted costs and normative costs is expected for one or more subgroups defined by the S-type adjusters:

where

$$\begin{cases} \Delta_t^s > 0: \textit{overcompensation} \\ \Delta_t^s < 0: \textit{undercompensation} \end{cases}$$

$$(2.6)\quad \Delta_t^s = \frac{1}{n_j(s)} \sum_{i \in I_j(s)} \hat{y}_{it} - \frac{1}{n_j(s)} \sum_{i \in I_j(s)} y_{it}^{NORM}$$

If the S-type adjuster takes on non-discrete values, then the subgroups are defined by the quantiles of the distribution of these non-discrete values in this study.

The performance of REF models in terms of their ability to generate risk-adjusted premium subsidies as intended by the sponsor can then be determined by the calculation of the weighted average of the (absolute values of) Δ_t^s over all subgroups defined by the S-type adjusters and comparing it against the value that would result in the absence of any risk-adjusted premium subsidies, where the weights are the population sizes of these respective subgroups. More precisely, the performance of a REF model is calculated as the percentage change in the weighted average of the (absolute values of) Δ_t^s relative to its value under the hypothesis of a complete absence of risk-adjusted premium subsidies in the health insurance market. Note that in the situation without any risk-adjusted premium subsidies only a constant term $\alpha_{0,t}$ is included in equation (2.1), which equals average observed costs for all insured people after estimation following equation (2.2).[32]

2.3.3 Aligning the REF weights with normative costs

The REF equation generates effective risk-adjusted premium subsidies if REF predicted costs equal normative costs for every subgroup defined by the S-type adjusters. In that case, there is no need to change the set of REF weights from equation (2.2). However, in this study it is assumed that the criterion of feasibility restricts the set of potential REF adjusters severely and therefore the risk-adjusted premium subsidies will deviate from the intentions of the sponsor. If REF predicted costs differ from normative costs for some subgroup defined by an S-type adjuster, then it is not obvious how to improve the REF equation in this respect. Indeed,

32. Quadratic of even higher powers can be taken of the deviations from normative costs instead of the absolute values used to calculate the performance indicator in this study. However, taking absolute values is most relevant in our opinion, because then this performance indicator has the same unit of measurement as the premium rates.

normative costs Y_t^{NORM} from equation (2.4) can not be computed for the total population of insured, because it is financially and logistically impossible to conduct a health survey among the total population, let alone on a continuous basis.

Although the premiums subsidies can not be improved by reducing the gap between REF predicted costs and normative costs for the subgroups defined by the set of the S-type adjusters, a reduction of the gap for the subgroups defined by the REF adjusters can be achieved. Such a reduction is desired because this gap must be attributed to N-type risk factors and risk-adjusted premium subsidies are actually distributed among subgroups defined by these REF adjusters in practice.

By definition, imperfect REF adjusters capture cost variation caused by N-type risk factors which is reflected as biased REF weights. As a solution to remove the bias in the estimated REF weights of these imperfect REF adjusters, Schokkaert, Dhaene and Van de Voorde (1998) and Schokkaert and Van de Voorde (2000, 2004) advocate the procedure to include variables during the estimation phase that capture the N-type cost variation and exclude them when calculating the premium subsidies. In this case, equation (2.1) is treated as a structural form equation in which the estimated REF weights suffer from omitted variables bias. This means that measures of both S-type and N-type risk factors are now included in equation (2.1), as described in equation (2.1'):

$$(2.1') \quad Y_t = \omega_{0,t} + \sum_{j=1}^{J} \omega_{j,t} * X_{j,t-1} + \sum_{k=1}^{K} \delta_{k,t} * Z_{k,t-1} + \kappa_t$$

where Y_t are the health care costs observed in year t, $X_{j,t-1}$ is the j^{th} REF adjuster observed in year t-1, j=1,...,J, $Z_{k,t-1}$ is the k^{th} measure of the N-type risk factors as observed in year t-1, k=1,..,K, $\omega_{j,t}$ and $\delta_{k,t}$ are unknown coefficients to be estimated and κ_t is an independent and identically distributed error term in year t. In order to derive an estimate of normative costs under this approach, the values of the $Z_{k,t-1}$ variables are set equal to some level desired by (or: 'acceptable' to) the sponsor. In practice, this desired level often equals the overall sample mean for each and every subgroup of enrollees, as is described by the following equation:

$$(2.2') \quad \hat{Y}_t = \hat{\omega}_{0,t} + \sum_{j=1}^{J} \hat{\omega}_{j,t} * X_{j,t-1} + \sum_{k=1}^{K} \hat{\delta}_{k,t} * \bar{Z}_{k,t-1}$$

Note that $\hat{\alpha}_{0,t} \neq \hat{\omega}_{0,t} + \sum_{k=1}^{K} \hat{\delta}_{k,t} * \bar{Z}_{k,t-1}$ and $\hat{\alpha}_{k,t} \neq \hat{\omega}_{k,t}$ for all k=1,..,K if the weights corresponding to the variables $Z_{k,t-1}$ are non-zero. The identity between (overall) average REF predicted costs and observed costs presented in the previous subsection only holds if the values of the $Z_{k,t-1}$ variables are set equal to the overall sample mean for each and every subgroup of enrollees.

Deviations from normative costs which can be attributed to N-type risk factors should be separated from the REF weights. This may be achieved by deriving REF predicted costs in an omitted variables bias context, as described by equation (2.2'). In order to determine the extent to which this procedure removes the bias from the REF weights, for each subgroup defined by the REF adjusters Δ_t^z from equation (2.5) given $\hat{Y}_t$ derived from equation (2.2') must be compared to Δ_t^z from equation (2.5) with REF predicted costs $\hat{Y}_t$ as derived from equation (2.2).

An alternative procedure to adjust the estimated REF weights $\hat{\alpha}_0$, $\hat{\alpha}_j$, $j=1,\ldots,J$ is advocated in this study. This procedure essentially boils down to regressing normative costs Y_t^{NORM} instead of observed costs Y_t on the limited set of REF adjusters. More specifically, the first step in this procedure is to estimate the following equation for the subsample of survey respondents:

$$(2.7)\quad Y_t^{NORM} = \gamma_{0,t} + \sum_{j=1}^{J} \gamma_{j,t} * X_{j,t-1} + \nu_t$$

with normative costs Y_t^{NORM} from equation (2.4) as the dependent variable. Notice that the set of risk adjusters $\{X_{j,t}, j=1,\ldots,J\}$ is exactly the same as the set of REF adjusters described in equations (2.1) and (2.2). Estimation of equation (2.7) then generates an estimate of normative costs $\hat{Y}_t^{NORM}$ that can be seen as an alternative to the estimate of normative costs that results from the REF model, i.e. REF predicted costs $\hat{Y}_t$. $\hat{Y}_t^{NORM}$ more closely resembles normative costs the larger the variance explained in this regression:[33]

$$(2.8)\quad \hat{Y}_t^{NORM} = \hat{\gamma}_{0,t} + \sum_{j=1}^{J} \hat{\gamma}_{j,t} * X_{j,t-1}$$

By construction, the so-called adjusted REF weight $\hat{\gamma}_j$ is an estimate of the marginal effect of the REF adjuster $X_{j,t-1}$ on normative costs Y_t^{NORM}, whereas the original REF weight $\hat{\alpha}_j$ is an estimate of the marginal effect of the REF adjuster $X_{j,t-1}$ on observed costs Y_t. In terms of Figure 2.2, the adjusted REF weight $\hat{\gamma}_j$ equals the slope of the true relationship between normative costs and the REF adjusters and the original REF weight $\hat{\alpha}_j$ equals the slope of the observed relationship. Therefore, if for some $j \in \{1,\ldots,J\}$ the difference between these marginal effects is non-zero, $\hat{\alpha}_j - \hat{\gamma}_j$ is an estimate of the marginal effect of the REF adjuster $X_{j,t-1}$ on observed costs for which the sponsor does not desire cross-subsidization. This

33. The estimated percentage of explained variance (R^2) in equation (2.8) is expected to be much larger than in equation (2.4), because at the individual level the variance of the dependent variable in equation (2.7) is much smaller than in equation (2.3).

will be the case if the REF adjusters define other subgroups of insured than do the S-type adjusters.

As the ordinary least squares technique is used to estimate the regression coefficients in equation (2.7), it follows that the estimate of normative costs $\hat{Y}_t^{NORM}$ is equal to normative costs Y_t^{NORM} for subgroups of insured defined by the REF adjusters on average:

$$\frac{1}{n_j(x)} \sum_{i \in I_j(x)} \hat{y}_{it}^{NORM} \equiv \frac{1}{n_j(x)} \sum_{i \in I_j(x)} y_{it}^{NORM}$$

Therefore, at least for the subgroups defined in this way, compensation for N-type cost variation can be avoided if the set of adjusted REF weights $\hat{\gamma}_0$, $\hat{\gamma}_j$, $j=1,\ldots,J$ is used instead of the original REF weights $\hat{\alpha}_0$, $\hat{\alpha}_j$, $j=1,\ldots,J$ in practice. Given the zero deviation of $\hat{Y}_t^{NORM}$ from Y_t^{NORM}, the extent to which this is indeed the case is equal to the deviation of $\hat{Y}_t$ from Y_t^{NORM}, where $\hat{Y}_t$ equals REF predicted costs from equation (2.2) for individual *i*. The conclusion is that the risk-adjusted premium subsidies for the total population of insured people should be based on the set of adjusted REF weights from equation (2.8) instead of the unadjusted REF weights from the REF equation (2.2).

It should be noted that, for subgroups defined by the REF adjusters, REF predicted costs are not necessarily equal to observed costs on average if the adjusted REF weights are applied instead of the original REF weights:

$$\frac{1}{n_j(x)} \sum_{i \in I_j(x)} \hat{y}_{it}^{NORM} \neq \frac{1}{n_j(x)} \sum_{i \in I_j(x)} y_{it}$$

This notion is in line with the suggestion in some recent theoretical papers that optimal risk adjustment does not generally require the capitation payments to equal average costs for the subgroups defined by the REF adjusters (Ellis, 1998; Frank et al., 2000; Glazer and McGuire, 2000, 2002; Sappington and Lewis, 1999).[34]

Of course, for subgroups other than those defined by the REF adjusters, normative costs will not be equal to REF predicted costs if adjusted REF weights are applied. Indeed, inequality will still hold for the subgroups defined by the S-type adjusters. In the following subsection, alternative specifications of the REF model are discussed as a way to improve the risk-adjusted premium subsidies for these subgroups.

34. At the moment, there is a rather large gap between this theoretical literature and the empirical work on the field of risk equalization. Such a gap does not exist under the approach developed in this study, however, given the direct relationship between equation (2.8) and equation (2.2).

2.3.4 Testing alternative specifications of the REF equation

An adjustment of the REF weights aligns REF predicted costs with normative costs for the subgroups defined by the REF adjusters. However, for the subgroups defined by the S-type adjusters, equality between REF predicted costs and normative costs only holds if the REF adjusters would form the basis of equation (2.4), which is not the case by construction. Therefore, the risk-adjusted premium subsidies can only be improved by an alternative specification of the REF equation.

For example, equation (2.1) may be estimated with an alternative set of risk adjusters. In Chapter 7 an empirical illustration of this strategy is given. The risk adjusters that are added, are all retrieved from the automated computer systems of the Dutch sickness fund Agis Health Insurance. Therefore, in principle, these can be made available from other Dutch insurers as well. In other words, the alternative specification is also feasible in Dutch practice.[35]

Another strategy to improve the risk-adjusted premium subsidies may be ex-post outlier risk sharing as a supplement to incomplete and/or imperfect REF adjusters. Outlier risk sharing may reduce the gap between $\hat{Y}_t$ and Y_t^{NORM}, with REF predicted costs $\hat{Y}_t$ redefined such that for each insured it includes the net effect of the contribution to and reimbursement from the risk sharing pool. However, this approach introduces a tradeoff between the improvement of the risk-adjusted premium subsidies and a reduction of the incentives for efficiency. A reduction of the incentives for efficiency occurs as a consequence of the ex-post character of this construct. Note that a combination of outlier risk sharing and proportional risk sharing is currently applied in Dutch practice.

As a third and final illustration, an alternative functional specification of equation (2.1) may also improve the risk-adjusted premium subsidies across the insured. Such alternative specification may be effective because REF equations are usually estimated by ordinary least squares instead of the cell-based approach, thereby often ignoring possible interactions between the REF adjusters. Furthermore, health care costs typically do not follow a normal distribution: health care costs tend to have a mode at zero costs and a distribution with a long, heavy right tail. In particular, variances of health care costs are usually not the same for every subgroup defined by the REF adjusters. As a consequence, the ordinary least squares estimates of the REF weights are not (asymptotically) efficient. This means that there exist other non-linear unbiased estimators of the REF weights which have

35. The scope of information can in principle be extended to data obtained from supplementary health insurance policies also. However, this information can not be used to quantify health status indicators to be included as REF adjusters in practice, as these are non-uniform policies across insurers.

smaller sampling variances. Therefore, in Chapter 7 both the multiplicativity and homoscedasticity assumptions are tested for the 2004 Dutch REF model.

In order to determine the contribution of each alternative specification of the REF model to the improvement of the risk-adjusted premium subsidies, the Δ_t^s is calculated according to equation (2.5) for all subgroups defined by the S-type adjusters. Again, the performance then equals the percentage change in the weighted average of this Δ_t^s relative to the weighted average of the Δ_t^s in the situation of a complete absence of risk-adjusted premium subsidies in the health insurance market.

2.3.5 Additional regulations for improving the subsidies

If the REF adjusters are incomplete measures of the S-type risk factors, then the risk-adjusted premium subsidies deviate from those intended by the sponsor. The sponsor may then decide to regulate the premium rates in order to create (implicit) subsidies across subgroups for uncaptured S-type cost variation. Rate restrictions can take several forms: rate-banding (by risk class), a ban on certain rating factors, and community rating (by risk class). Ideally, rate regulation creates (implicit) cross-subsidies for S-type cost variation alone.

Although rate regulation may create (implicit) cross-subsidies for S-type cost variation, it also creates predictable profits and losses for subgroups of insured people defined by the corresponding rating factor. Premium rate regulation therefore creates incentives for selection with adverse effects on quality of care, affordability and efficiency. Ideally, incentives for selection are avoided.

Predictable profits and losses are defined as the difference between REF predicted costs and observed costs, on average for a subgroup of insured people:[36]

$$\begin{cases} \bar{\hat{Y}}_t > \bar{Y}_t : \textit{predictable profit} \\ \bar{\hat{Y}}_t < \bar{Y}_t : \textit{predictable loss} \end{cases}$$

According to the identity

$$(2.9) \quad \left(\frac{1}{n_Z}\sum_{i\in I_Z}\hat{y}_{it} - \frac{1}{n_Z}\sum_{i\in I_Z} y_{it}\right) = \left(\frac{1}{n_Z}\sum_{i\in I_Z}\hat{y}_{it} - \frac{1}{n_Z}\sum_{i\in I_Z} y_{it}^{NORM}\right) + (-\Delta_t^Z)$$

the predictable profits and losses can be separated into two effects: a compensation effect caused by incomplete and/or imperfect REF adjusters and a utilization effect $-\Delta_t^Z$ as defined by equation (2.5) for which the sponsor desires no cross-subsidization. Premium rate regulation induces cross-subsidies in line with the

36. Note that predictable profits and losses are non-existent for the subgroups defined by the REF adjusters by construction.

intention of the sponsor to the extent that the first term on the right-hand side of equation (2.9) dominates the predictable profits and losses created by this regulation. Of course, the tradeoff with the incentives for selection remains even if this term dominates completely.[37]

If the REF model already generates risk-adjusted premium subsidies as intended by the sponsor, then REF predicted costs and normative costs coincide and the first term at the right-hand side of equation (2.9) equals zero. Additional regulation such as premium rate restrictions is then redundant and only creates predictable profits and losses with respect to underutilization and overutilization by specific subgroups of insured people caused by N-type risk factors. In that case, the premium rate restrictions should be abolished in order to avoid the incentives for selection and their adverse effects. If the REF adjusters are incomplete, however, then there exists a tradeoff between the effectiveness of the risk-adjusted premium subsidies and the incentives for selection. This tradeoff can be made explicit under the approach developed in this study.

2.4 Conclusions

In a competitive health insurance market, risk-rated premiums may be extremely high for high-risk individuals. In order to safeguard affordability, cross-subsidies from low-risk to high-risk individuals may be distributed via a so-called Risk Equalization Fund (REF). Theoretically, the costs for which the sponsor desires cross-subsidization may be the so-called acceptable costs. As the level of acceptable costs of the benefits package is hard to determine in practice, usually the risk-adjusted premium subsidies are based on observed costs instead. However, although risk-rated premiums tend to capture *all* systematic cost variation in competitive health insurance markets, the sponsor may desire the risk-adjusted premium subsidies to only compensate for cost variation among subgroups of insured people that are defined by the so-called S-type risk factors (Schokkaert and Van de Voorde 2000). The costs caused by the S-type risk factors are called normative costs. In this study it is assumed that Dutch government considers age, sex and health status to be S-type risk factors.

In practice, proxy measures of the theoretical S-type risk factors are included in the REF equation, called REF adjusters. The choice of REF adjusters is usually

37. In general, non-existent predictable profits and losses may occur as a result of either a combination of overcompensation and overutilization or a combination of undercompensation and underutilization.

restricted because these have to satisfy the criteria of effectiveness of the risk-adjusted premium subsidies, appropriateness of incentives and feasibility. The set of REF adjusters may therefore appear to be incomplete and/or imperfect measures of the S-type risk factors. In conventional REF models the availability of health status measures at the individual level of enrollees is rather limited, because they should be obtained from administrative databases that contain the information for the total population of members. Therefore, the set of REF adjusters may be incomplete. Furthermore, the administrative variables may not exclusively reflect S-type risk factors, but also N-type risk factors (e.g. supply). As a consequence, the REF adjusters can also be seen as imperfect measures of the S-type risk factors. In the Dutch context, this latter problem holds with respect to the REF adjusters eligibility and region.

Although one of the most sophisticated risk equalization formula in the world is currently in use in the Netherlands, it is still an open question to what extent the risk-adjusted premium subsidies fully satisfy the criterion of effectiveness. The contribution of this study is the development and application of a procedure to test REF models for their effectiveness. The results of this test procedure lead to concrete suggestions for improvement of the REF model currently in use.

The test procedure starts with the collection of a broad array of health status measures at the individual level, by means of a health survey that is conducted among a subsample of insured people. For the survey respondents, a more precise measure of the risk factor health status is obtained in this way. Predicted costs that follow from this elaborate risk equalization formula are called the normative costs, and are supposed to be those costs that reflect the S-type risk factors as precisely as possible. The difference with the REF model is that for this subsample of enrollees there are no limitations of feasibility. The derivation of normative costs gives an answer to the first research question of this study, i.e. how to find the costs for which Dutch government desires risk-adjusted premium subsidies.

From the literature on risk adjusters it appears that the most promising measures of the S-type risk factor health status are self-reported measures of perceived health status, functional health status and chronic conditions. In order to guide the specific selection of these S-type adjusters, the conceptual model of Ruwaard and Kramers (1997) is applied. The health status risk adjusters that are included in the normative risk equalization equation are the eight SF-36 scales, OECD scores and the number of a list of specific chronic conditions. PCGs and DCGs are added to the normative equation as it is not necessarily observed in the data that more medical care is used with more severe problems (Newhouse 1989).

Under the assumption that there exists a gap between REF predicted costs and costs for the subgroups defined by the REF adjusters, the gap may be reduced by regressing costs instead of observed costs on the REF adjusters (i.e. estimate equation 2.5 instead of equation 2.1). This gives the weights that accurately reflect risk-adjusted premium subsidies as desired by the Dutch government. These so-called adjusted REF weights can be applied to the total population of members because it makes use of the REF adjusters which are available for all. In the second section of Chapter Six, the adjusted REF weights are determined and a comparison is made between costs and predicted costs using these adjusted REF weights. These results are confronted with those from the omitted variables bias procedure to adjust REF weights, as proposed by Schokkaert, Dhaene and Van de Voorde (1998) and Schokkaert and Van de Voorde (2000, 2004). Based on these exercises, an answer can be found to the second research question in the context of the 2004 Dutch REF model.

In order to improve the risk-adjusted premium subsidies for the subgroups defined by the S-type adjusters, possible solutions are to find additional risk adjusters to include in the REF model, to apply ex-post risk sharing as a supplement to the incomplete and/or imperfect REF adjusters, or to maintain the same set of REF adjusters but use an alternative specification of the functional form and error distribution. In all these cases, a comparison between REF predicted costs and normative costs for the subgroups defined by the S-type adjusters determines the extent to which there is an improvement in the amount of induced cross-subsidization, and thus the success of the chosen variant. The results of these exercises are presented in Chapter Seven and give an answer to the third research question.

In Chapter Eight, the tradeoff between the effectiveness of the cross-subsidies and incentives for selection – induced by premium rate restrictions – is made explicit under the approach developed in this study. This exercise is done for subgroups defined by the (incomplete and imperfect) REF adjusters, for subgroups defined by self-reported prior medical utilization, self-reported health status, diseases and conditions, for the subgroups defined by the number of years that survey respondents belong to the top 25% of total expenses within each year prior to 2002, and for subgroups defined by the twelve Dutch provinces.

Chapter 3

Data

In order to find an answer to the research questions of this study, an instrument is needed to describe health status of the enrollees that belong to the Agis sickness fund population. Health status in the 2004 Dutch REF model is captured by the REF risk adjusters derived from the administrations of Dutch sickness funds. Section 3.1 describes the panel dataset 1999-2002 derived from the Agis sickness fund administration.

A more direct operationalization of health status is needed for the application of the normative test procedure proposed in this study. Section 3.2 describes the motivations for the selection of the survey questions, compares Agis prevalences with national figures in order to determine statistical representativeness, and describes the choice of respondents as eligible for analysis in this study.

In Section 3.3 the data are described that are obtained from other sources than the administration and mail survey.

3.1 Agis administrative data 1997-2002

The panel dataset 1999-2002 contains administrative data on sickness fund enrollees of "Agis Health Insurance", a June 1999 merger of "Anova Insurance", "Anoz Insurance", and "ZAO Health Insurance". With about 1.6 out of about 10.0 million sickness fund members in the Netherlands in 2002, Agis had the largest market share in the Dutch sickness fund market at the time. The panel dataset only contains records that are included in the national dataset on which the REF model is estimated by the CVZ agency of the Dutch Ministry of Health, Welfare and Sports.[38],[39]

The yearly administrative datasets consist of annual per-person health care expenses and the member characteristics gender, age, ZIP-code, membership length, and membership eligibility. Health care expenses at the individual level include general practitioner (GP) care, pharmaceutics, both inpatient and outpatient specialist care, dental care, obstetrics, inpatient room and board, paramedic care (physiotherapy, César/Mensendieck, speech and ergo therapy), medical devices, sick-transport, maternity care, and accountable costs of innovative care arrangements. The pharmaceutical drugs expenses prescribed by physicians from within the hospital are not included, only the drugs expenses that are delivered out of the

38. With the exception of the Anoz 2000 data.

39. The Dutch statistical office for health insurers called Vektis collects, screens, and corrects the data under the authority of the Dutch Ministry of Health. Eligibility rules are set up in cooperation with health insurers.

hospital by the pharmacist. The GP expenses are not related to medical consumption, as GPs receive capitation payments for their services at the time. All data on expenditures refer to actual charges.

The 2001 claims data are available for all the health care services listed above. These claims data are not yet aggregated to the person level. First of all, they are used to construct the so-called pharmacy-cost groups that are included as REF risk adjusters in the 2004 REF model. Furthermore, the 2001 paramedic, medical devices and mentally oriented drugs claims data are used to construct new sets of risk adjusters to apply the normative test procedure to an alternative specification of the REF model in Section 7.1.

3.2 Agis Health Survey 2001

Subjective measurement of health status amongst a general population of Agis enrollees is judged most applicable for the purpose of this study. Although sampling objective medical information might seem preferable at first sight, there are also important drawbacks such that objective measurement is not the preferred option for the purpose of this study. For example, medical files of GPs will not always reveal a complete medical profile of each and every enrollee that participates in this study. For enrollees not registered at any GP's practice this is most obvious. Furthermore, Mackenbach, Loomanm, and Van der Meer (1996) note that general practitioners do not always have an accurate idea of the diagnosis of their patients, especially if the patient is in fact under treatment of the medical specialist instead of the general practitioner, or under no treatment at all.

Another drawback of relying on objective measurement of health status is that for those that had a GP contact, the process of data collection is very costly and time consuming or even impossible. For example, in the first measurement round in 1991 of the longitudinal Van der Meer et al. (1996) study, medical information about conditions and GP treatment could be obtained for only 38% of the 3,970 people that had been identified to suffer from one specific disease[40]. It is expected that collecting medical information implies even more effort if a more general population is targeted.

On the other hand, with respect to subjective measurement of health status, Mackenbach, Loomanm, and Van der Meer (1996) note that the information

40. These people had been identified beforehand to suffer from heart and cardiovascular diseases, asthma/COPD, back pain or diabetes mellitus and lived in the city of Eindhoven or one of its 17 surrounding municipalities. Permission to ask their GP after their condition and treatment was given by 72%.

sampled from surveys is not insensitive to the perceptions of individuals. For example, self-reported diseases as indicated by a very common survey checklist showed lower prevalences than extracted from this medical information, with the exception of diabetes.[41] However, they are not able to choose between objective measurement and subjective measurement as the golden standard of health status measurement.

In conclusion, measuring health status based on medical files at the general practitioner's is not the way to go for our study. It is decided that health status will be measured subjectively by means of a health survey, which seems to be a more cost effective approach especially given the time schedule of this study.[42]

A description of the choices for the questions to be included into the survey is presented in Section 3.2.1. In Section 3.2.2 statistical representativeness of the sample of survey respondents is discussed. Section 3.2.3 describes the selection of respondents eligible for analysis in this study.

3.2.1 Questionnaire design

There are different data collection modes to choose from when conducting a health survey, e.g. the survey can be conducted by mail, by telephone, in-person, etcetera. For the purpose of this study, the mailing mode is chosen as the preferred mode of data collection.[43]

With respect to the selection of questions to incorporate in the Agis Health Survey 2001, a main starting point was the 1993 health survey conducted amongst enrollees of another Dutch sickness fund with about 0.4 million enrollees, called Zorg & Zekerheid (Z&Z). The Z&Z survey data are used extensively by Lamers (1997) and Van Barneveld (2000) in their research on risk adjustment models.[44]

Questions in the Z&Z Health Survey 1993 are mainly drawn from the national survey Permanent Research on Living Conditions (POLS), as conducted yearly by

---

41. According to Van der Meer et al. (1996) it may be the case that people with diabetes are more adequately informed about their disease and/or that this diagnoses is always known with the general practitioner.

42. As an alternative to gathering information at the general practitioner's, the number of enrollees surveyed may be increased to reduce measurement error. Validation of reported diseases may be based on Agis claims data (this exercise is not pursued in this study).

43. See Appendix A3.1 for a motivation for the choice of a postal survey as the preferred administration mode to collect the subjective health status data, and a description of the data collection process, starting with drawing a stratified sample from the total Agis population, up to receiving the scanned data from the survey vendor.

44. Appendix A3.2 discusses the results of a separate regression analysis in order to determine the predictive power of the Z&Z survey variables with respect to 1994 Dutch health care expenditures.

Statistics Netherlands (CBS). The CBS is the Dutch agency that is responsible for the (official) national statistics. POLS contains questions on short-term disabilities, long term health problems and chronic diseases, quality of life, lifestyle, social and physical environment. Institutionalized people are excluded from the sample.

Survey questions included in the Agis Health Survey 2001 are selected in joint cooperation with one of the members of the so-called Working Group "Revision POLS-Health Survey 1999".[45] This working group advised on a major revision of the Dutch national health survey, the final report being published by Van den Berg and Van der Wulp (1999). The questions to be included into the Agis Health Survey both had to be relevant in the context of the current study and in line with the advise of this working group in order to generate results that can be used nation wide.

The Agis Health Survey 2001 consists of four main sections: "Health status", "Sickness", "Use of care", and "Background characteristics". The survey questions chosen for each section are described in more detail hereafter.

Survey section "Health status"

The first section on "Health status" (**questions 1-11**) is formed by the Dutch translation of the SF-36 questionnaire form (see Aaronson et al., 1998). The SF-36 is a 36-item instrument for measuring health status and outcomes from the patient's point of view and was designed for use in clinical practice and research, health policy evaluations, and general population surveys. The most common SF-36 dimensions of health status are physical, mental and social functioning. The SF-36 measures the following eight health concepts:

- limitations in physical activities because of health problems;
- limitations in usual role activities because of physical health problems;
- bodily pain;
- general health perceptions;
- vitality (energy and fatigue);
- limitations in social activities because of physical or emotional problems;
- limitations in usual role activities because of emotional problems; and
- mental health (psychological distress and well-being).

The SF-36 grew out of work on the Medical Outcome Study or RAND Health Insurance Experiment (Ware and Hays, 1988). A 36-item short-form was constructed out of a need to gather health status information on individuals who did not answer a longer form during the RAND Health Insurance Experiment. As documented in

45. We thank dr. M. Foets for her kind cooperation. Of course, the choices made in this study are our own responsibility.

more than 1600 publications, the original SF-36 proved useful in (1) monitoring general and specific populations, (2) comparing the burden of different diseases, (3) differentiating the health benefits produced by different treatments, and (4) screening individual patients.

In this study, a generic measurement instrument is chosen, as health status should be measured for a general population consisting of people with and people without (non-specific types of) diseases. Among the generic health status measurement instruments there exist the Sickness Impact Profile, Nottingham Health Profile, COOP/Wonca-charts, SF-36, EuroQol, and Health Utilities Index.[46],[47] The SF-36 (i.e. a Short-Form health survey with 36 items) is chosen for the purpose of our study for several reasons.

First of all, Van den Berg and Van der Wulp (1999) recommend to use the SF-36 in Dutch national health status surveys. As already mentioned above, the questions to be included into the Agis Health Survey should be in line with the advise of this working group in order to generate results that can be used nation wide.

Second, the SF-36 Health Survey is the most widely used health status survey in the world, as it is translated into more than 40 languages and administered to millions of people worldwide. Its psychometric properties in terms of validity and reliability are well-documented and results are also comparable between countries. In a bibliographic study of the growth of "quality of life" measures, Garratt et al. (2002) judge the SF-36 to be the most widely evaluated generic patient assessed health outcome measure.

A third reason for choosing the SF-36 is that this is in line with the choice of the Centers for Medicare and Medicaid Services (CMS) to include the SF-36 as the main instrument for collecting outcomes data from Medicare beneficiaries in the Health Outcomes Survey (HOS). The HOS is initiated in 1996, from a recognized need to monitor the performance of managed care plans. It is the first national survey to measure the health status of Medicare beneficiaries enrolled in managed care. The HOS program seeks to gather valid and reliable health status data for use in quality improvement activities, public reporting, health plan accountability and improving health outcomes based on competition. The HOS data are also a major constituent of the Health Plan Employer Data and Information Set (HEDIS). For a recent evaluation of the HOS program, see Jones, Jones, and Miller (2004). For an introduction to the HEDIS program, see NCQA (2005).

---

46. For an overview of widely used general health surveys and the concepts they measure, see e.g. Iezzoni (2003, p. 61) or Ware (1995, p. 330).

47. Note that in the context of this study, an instrument is needed to describe health status differences only. Valuations of health status differences are not needed.

The first two SF-36 items (**questions 1 and 2**) are about self-assessed health in general, now and as compared with the situation in the past. These questions are similar to the first two questions in the Z&Z Health Survey 1993, although the wording of the question and that of its response categories is somewhat different. Furthermore, the reference period in the second question differs (last year versus five years ago, respectively). The SF-36 convention is followed in the Agis Health Survey 2001.

The third Z&Z survey question to derive the so-called VOEG score has not been included in the Agis Health Survey 2001. According to Van den Berg and Van der Wulp (1999), this question is old-fashioned and barely used. The Affect Balance Scale (ABS) that was included in the Z&Z survey is also dropped from the Agis survey for the same reason. Furthermore, they state that although the VOEG score is intended to measure mental health, this instrument predominantly measures physical limitations. The corresponding SF-36 items capture psychological distress as well.

Survey section "Sickness"

In the section on sickness and diseases, the Z&Z question on staying home in bed is included in the Agis survey (**question 12**). Also the same reference period of 14 days is applied here, as from the Z&Z survey it appeared that the group of people that stayed 14 days in bed did not appear to be too large to be distinctive (only 7% of persons have stayed home for 8 days or more). From Lamers (2000) it appeared that this variable has a significant effect on health care expenses, given some variants of the Dutch REF model.

Seven items on acute complaints are asked for in this mail survey (**question 13**). However, they are not included in the normative regression equation because they are less relevant in prospective REF models.

The Z&Z list with long-term diseases is also included in the Agis health survey, with some refinements in accordance with the advise from Van den Berg and Van der Wulp (1999). This amounts to a list of twenty long-term diseases (incl. an "other long-term illness or disease" category), of which five diseases form the basis for a count variable to be included in the normative regression equation (2.3) as a risk adjuster: diabetes mellitus (type I and II), stroke/brain haemorrhage/ infarction, myocardial infarction, other serious heart disease, and some type of (malignant) cancer (**question 14**). These diseases are only taken into account in the normative risk adjuster if the respondent still has complaints or is still under treatment.

Two items on fear and depression are included in the Agis Health Survey 2001 in order to measure psychological problems (**question 15**). As mental health is

supposed to be measured already adequately by the SF-36 items, in the empirical analysis these fear and distress items are only used in order to evaluate outcomes for constructed subgroups of enrollees. Van den Berg and Van der Wulp (1999) note that other psychological problems are too difficult to measure.

The remaining list of fifteen out of twenty long term diseases are only used to form subgroups of enrollees for which REF predicted costs are compared to normative costs (**question 16**). The reason for not including them in the normative equation is that they are already captured (2.3) via some pharmacy-cost group (asthma/COPD and chronic joint inflammation), they may go by (serious/persistent back problem, injuries of neck, shoulder, elbow, wrist, and hand), they are disorders instead of diseases (dizziness when falling down, intestinal obstructions, and urinary incontinence), or the severity of the disease for a given subgroup of respondents is simply more diffuse than the five long-term diseases that are included in the normative equation (migraine or serious headache regularly, hypertension, vascular constriction, psoriasis, chronic dermatitis, osteoarthritis).

Survey section "Use of care"

In this section questions on the use of health care are posed. Of course, many of the information needed can be drawn from the Agis administrative claims data as well. However, apart from being able to cross-check these claims data, not all contacts with health care providers and institutions can be derived from the claims files. For example, the contacts with the GP are not recorded, as GPs receive capitated payments for their services.

In the section "Use of care" most Z&Z questions are also included in the Agis Health Survey 2001, except for the questions on expectations with respect to future health status. From a separate analysis, these questions appeared to have little predictive value with respect to health care expenses (Lamers 2000). The OECD question on functional limitations concludes this survey section, although this question in fact belongs to the "Health status" section (see the discussion above).

The Z&Z question on contacts with the general practitioner have a recall period of 12 months (**question 17**). In case of the subquestion on the number of contacts, the recall period is 2 months. Note that in order to be able to transform this latter number into a yearly number, it would be necessary to know the date when the survey was completed. However, only the date when the survey form is received by the survey vendor will be known. As in this study not the exact number of contacts but only the intensity of contacts is needed, such a transformation is not needed for the purpose of this study.

The recall period with respect to the specialist contacts is 12 months (question 18a). From the the POLS survey it appeared that a recall period of 2 months led to very low prevalences, therefore in the Z&Z survey as well as the Agis survey a recall period of 12 months is applied. Furthermore, specialist (or co-assistant) contacts during hospital stays are excluded, as daily contacts may increase these numbers without there being any relationship with patient severity. In contrast to the Z&Z survey, the type of specialists is not asked for in the Agis survey.[48]

The Z&Z survey questions on expected specialist and hospital/clinical contacts are not included in the Agis survey. From Lamers (1997) it follows that the predictive power of stated expectations for future expenditures of specialist or hospital contacts is negligible. The question on overnight hospital or clinical stays is posed in the Agis survey (**question 18b**).

The questions on dental plates and dental care may reflect health status and/or socio-economic status: the higher your income, the longer you may conserve your own teeth. However, from a separate (non-published) analysis in the context of the Lamers (1995) study it appeared that these variables only had an impact with respect to supplementary insurance. As supplementary health insurance is out of the scope of this study, questions on dental plates and dental care are not included in the Agis survey.

Questions on paramedic (**question 19**), psycho-social (**question 20**) and alternative care providers (**question 21**) are included in the Agis survey. Furthermore, questions on home care are included in order to determine how needy people are for daily assistance.

There are questions on the amount of home care that is being used, covered by the insurance policy (**question 22**) or supplied by family and friends (**question 23**). This information may be used in order to construct subgroups of enrollees to test REF predicted costs against normative costs.

Use of prescribed and non-prescribed pharmaceutical drugs is asked for in the Agis survey just as in the Z&Z survey (**question 24**). However, because the administrative claims of prescription drugs are available from the automated systems, the type of prescription drugs has not been asked for in the Agis survey.

Z&Z survey questions on supplementary insurance and mandatory deductibles for prescription drugs are not included in the Agis survey, because these are out of the scope of this study.

Van den Berg and Van der Wulp (1999) argue that the SF-36 does not include all OECD and ADL items and therefore recommend the OECD question on long-term

48. Except help from a psychiatrist, which was added to the question on psychological and/or social care provision.

functional limitations, as well as the ADL question on activities of daily living for which people need assistance.[49] In the Agis Health Survey only the OECD question is included (**question 25**). The reason for not including the ADL items is that from Van den Berg and Van der Wulp (1999) and HEDIS (2003) it appears that these are more appropriate when focusing on people with physical disabilities and/or the elderly population, whereas the OECD items are more generically applicable.[50] A choice had to be made between the two questions, because only a limited amount of questions could be included.[51] Functional health status as measured by the number of OECD limitations is included in the normative regression equation (2.3).

Survey section "Background characteristics"

The fourth and final section concerns the background characteristics, in order to be able to determine to what extent socio-economic differences explain differences in health status and/or health care use. Questions on gender (**question 26**) and date of birth (**question 30**) are asked for validation reasons, as these are also recorded in the available data on the insurance policy.

In her research, Lamers (1997) has included questions on the ethnicity of respondents and their parents in order to determine predictive ratio's for these subgroups. From her Table 9.1 it appears that there is underutilization, i.e. actual costs of respondents with parents born outside the Netherlands are significantly lower than REF predicted costs if based on a demographic model. However, whether REF predicted costs adequately captures normative costs for these immigrants is another type of question and the answer to that has yet to be determined. In order to find an answer, questions on the ethnicity of respondents are included in the Agis survey (**questions 27-29**).

Weight and height are asked for in order to be able to determine underweight, overweight, or obesitas (**questions 31-32**). It should be noted that collection of this type of information by mail survey may be problematic because this is not a

---

49. In the Netherlands, ADL questions are used to determine whether an individual qualifies for admission to a nursing home or for use of paid home care.

50. Three out of seven ADL items are appropriate for the 12+ population as well, but it is decided that they are not to be included in the Agis Health Survey nonetheless. These three ADL items are "Getting in and out of bed", "Go up and down the stairs", and "Getting in or out of chairs".

51. In the HOS, the ADL items are included because they allowed lower levels of physical functioning to be measured better than with the SF-36 PCS scale alone (HEDIS 2003). Note that the HOS is conducted for the Medicare beneficiaries, i.e. for people being 65 or older and beneficiaries under the age of 65 with disabilities.

closed-form question. In addition, the hand-written answers have to be processed digitally. Therefore, a relatively high partial nonresponse is expected.

There are questions on education (**question 33**), on marital status (**question 34**), whether the respondent is the main source of income or not (**question 35**), and on household size (**question 36**) included in the Agis survey.

The last question in the Agis survey is on income (**question 37**). In the Z&Z survey, there were questions on employment status instead of income. However, employment status may be based on information on insurance eligibility that is available from the insurance policy data.

From Van den Berg and Van der Wulp (1999) it appears that the lifestyle questions on smoking, drinking and physical exercise should be extensive if one wants to have valid measurements. Because of limited availability of survey space, these lifestyle questions are not included in the Agis survey.

3.2.2 Statistical representativeness

In this section the national representativeness of the sample of Agis respondents to the Agis Health Survey 2001 is determined with respect to their health care contacts, health status, lifestyle, and some background characteristics.[52] A comparison is made with Dutch national figures as derived from the POLS 2001 survey.[53] The figures derived from the Agis Health Survey 2001 are adjusted (by means of direct standardization) to the 2001 Dutch national age and gender distribution of the sickness funds population (averaged over 12 months), as published by CVZ (2001).[54],[55]

Table 3.1 shows that the self-reported contacts of Agis members with general practitioners, medical specialists, hospitals, physiotherapists, RIAGGs, and alternative care practitioners are more or less comparable to that of the Dutch sickness fund population. The only exception seems to be the RIAGG, but that discrepancy may be explained by a difference in reference periods.

Discrepancies between the Agis Health Survey 2001 and the CBS POLS 2001 survey are also present with respect to the pharmaceutical drugs. More specifically, it appears that a larger number of Agis enrollees used prescribed drugs in

52. A detailed description of data collection process can be found in Appendix A3.3. An overview of the self-reported variables used to test statistical representativeness can be found in Appendix A3.4.

53. The POLS figures can be found at the CBS Statline website (http://statline.cbs.nl/).

54. Age is classified in the categories 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85 and older.

55. From the response and nonresponse analysis presented in Appendix A3.5 it turns out that no standardization for selective nonresponse is needed in this study.

Table 3.1: Use of medical services by sickness fund enrollees (weighted for Agis sample vs Agis population differences). CBS figures reflect the population of Dutch sickness funds enrollees only.

Variable [a]	Agis HS 2001	Agis HS 2001, standardized [b]	CBS POLS 2001 ZFW-NL [c]
% Enrollees with GP contact	77.2	75.9	78.0
% Enrollees with specialist contact	41.5	40.0	39.2
% Enrollees with hospitalization	8.3	7.9	6.3
% Enrollees with paramedic contact	18.2	17.4	17.2
% Enrollees with RIAGG contact	8.1	8.0	2.1
% Enrollees with alternative care contact	11.0	10.4	12.3
Pharmaceutical drugs			
% enrollees with prescribed drugs	48.7	45.7	37.8
% enrollees with non-prescribed drugs	27.4	26.8	35.8

[a] In Appendix A3.2 a detailed description is given of which survey questions are used for this table and how the subgroups are derived from the answers to these questions.
[b] The weighted figures constructed from the Agis HS 2001 are adjusted (direct standardization) to the 2001 Dutch national age and gender distribution of the sickness funds population (averaged over 12 months), as published by CVZ (2001). Age is classified in the categories 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85 and older.
[c] The CBS Statline figures are based on the subset of sickness fund members that participated in the 2001 POLS survey. Roughly two-thirds (i.e. about 6,451 cases) of the 9,676 respondents to the POLS survey (Botterweck et al., 2003) therefore constitute the Dutch national figures from the POLS survey as presented in this table, as this is the share of the sickness fund members in the total population. Furthermore, note that the Agis HS 2001 figures only hold for Agis members of 16 years and older. Therefore, under the assumption that medical consumption of people below 16 is lower than medical consumption of those above 16 years of age, the standardized Agis HS 2001 figures in this table will be biased upward when compared to those for the total Dutch population of sickness fund members.

the 14 days before the survey was filled out than in the Dutch population. On the other hand, the percentage of Agis enrollees that used non-prescribed drugs is substantially lower than for the Dutch sickness fund population.

In Table 3.2 self-reported diseases and lifestyle are compared between the Agis Health Survey 2001 and the Dutch national CBS POLS 2001 survey. In general, self-reported long-term diseases appear to be more present amongst the Agis enrollees than in overall Dutch society. Physical health is somewhat above national average, although functional limitations according to the OECD information indicates otherwise.[56] Mental health appears to be somewhat below national average,

56. Note that the Agis population is more representative for the Dutch sickness fund population, as they include those members living in pensioner homes. This is supported from a comparison of the OECD scores with that of the Z&Z survey in Lamers (1995).

Table 3.2: Self-reported diseases, health status and lifestyle

Variable [a]	Agis HS 2001	Agis HS 2001, standardized [b]	CBS POLS 2001 [c]
Long-term diseases			
Diabetes mellitus (Type I and II)	4.5	4.0	3.4
Stroke, brain haemorrhage/infarction	2.6	2.4	1.5
Myocardial infarction	3.4	3.2	2.2
Other serious heart disease	2.2	2.1	1.4
Some type of (malignant) cancer	4.7	4.2	3.5
Migraine or serious headache regularly	21.8	21.4	16.5
Hypertension	14.7	13.3	8.8
Vascular constriction (stomach, legs)	4.2	3.9	1.7
Asthma, COPD	8.1	7.7	7.5
Psoriasis	1.7	1.6	1.7
Chronic dermatitis	5.1	5.1	4.0
Dizziness when falling down	6.5	6.5	3.0
Intestinal obstructions (> 3 months)	4.4	4.2	2.7
Urinary incontinence	7.5	6.8	4.5
Serious/persistent back problem	14.6	13.8	9.3
Osteoarthritis (hip/knees)	15.7	14.2	9.5
Chronic joint inflammation	6.1	5.3	3.9
Other serious/persistent injury (neck, shoulder)	14.8	13.8	9.7
Other serious/persistent injury (elbow, wrist, hand)	9.5	8.8	6.6
Other prolonged disease/disorder	11.8	11.3	8.3
Generic health status measure			
Physical Component Scale score	53.3	53.7	49.8
Mental Component Scale score	50.7	50.8	52.0
Psychological distress			
Fearful or afraid (for 2 months)	29.8	28.4	30.3
Downhearted or blue (for 2 months)	29.9	29.2	27.8
Either fearful/afraid or downhearted/ blue (for 2 months)	38.5	37.3	38.9
Functional limitations			
% enrollees with one or more OECD limitations	22.4	20.8	14.2
Number of OECD limitations per enrollee	0.4	0.4	0.2
Number of OECD limitations per enrollee with limitation	2.0	2.0	1.7
% enrollees with OECD auditive impairment	5.3	5.2	2.9
% enrollees with OECD visual impairment	8.1	7.5	4.3
% enrollees with OECD mobility impairment	16.1	14.9	9.8

Variable [a]	Agis HS 2001	Agis HS 2001, standardized [b]	CBS POLS 2001 [c]
Lifestyle			
Height	171.4	172.5	172.9
Weight	73.9	74.2	74.9
% enrollees with underweight = BMI less than 18.5	3.1	3.4	2.0
% enrollees with normal weight = BMI 18.5 - 24.9	51.6	53.1	52.5
% enrollees with overweight = BMI 25 - 29.9	45.3	43.5	45.5
% enrollees with obesitas = BMI of 30 or greater	12.4	11.6	10.8

[a] In Appendix A3.2 a detailed description is given of which survey questions are used for this table and how the subgroups are derived from the answers to these questions.
[b] The weighted figures constructed from the Agis HS 2001 are adjusted (direct standardization) to the 2001 Dutch national age and gender distribution of the sickness funds population (averaged over 12 months), as published by CVZ (2001). Age classified in the categories 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85 and older.
[c] The CBS Statline figures are based on the subset of sickness fund members that participated in the 2001 POLS survey. Roughly two-thirds of the 9,676 respondents to the POLS survey (Botterweck et al., 2003) therefore constitute the Dutch national figures from the POLS survey as presented in this table (i.e. about 6,451 cases), as this is the share of the sickness fund members in the total population. Long-term diseases are representative for the total Dutch population. Generic health status, psychological distress, and functional limitations is measured amongst those of 12 years or older, and height and weight for persons of 20 and above. Note that the Agis HS 2001 figures only hold for Agis members of 16 years and older.

although the figures with respect to psychological distress show a somewhat ambiguous picture.[57] Height and weight are similar to national figures, although in terms of the Quetelet index, the extremes of underweight and obesitas are more prevalent amongst Agis enrollees.

Table 3.3 shows that first and second generation immigrants are underrepresented in the Agis Health Survey 2001 compared to the total Dutch population. However, note that enrollees younger than 16 years of age are excluded from this study.

Table 3.4 presents the classification of respondents according to the highest level of successfully finished education. Enrollees with elementary and lower secondary education are overrepresented amongst Agis enrollees as compared to Dutch national figures. The main explanation for this difference is the fact that in order to become a sickness fund member in 2001, individual gross income has to be below € 29,813. From Lamers (1995) it appears that in the 1993 Z&Z sickness fund

57. National figures are derived from a Dutch version of the SF-12 questionnaire.

Table 3.3: First and second generation immigrants

Variable	Agis HS 2001	Agis HS 2001, standardized [a]	CBS 2001 [b]
Non-immigrants	91.9%	91.9%	81.6%
First generation immigrants	5.7%	5.6%	9.6%
Second generation immigrants	2.4%	2.5%	8.8%
Total	100.0%	100.0%	100.0%

[a] The weighted figures constructed from the Agis HS 2001 are adjusted (direct standardization) to the 2001 Dutch national age and gender distribution of the sickness funds population (averaged over 12 months), as published by CVZ (2001). Age classified in the categories 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85 and older.
[b] The CBS Statline figures are based on the population estimates of the total Dutch population, dated January 1, 2002. Note that those figures also cover non-sickness fund members (about one third of the total population) and people younger than 16.

Table 3.4: Highest level of education, enrollees 16-64 years of age

Variable	Agis HS 2001	Agis HS 2001, standardized [a]	CBS 2001 [b]
Elementary education (incl. not finished)	20.6%	20.2%	12.9%
Lower secondary education (Lbo/Mavo/Vmbo)	38.8%	38.1%	25.5%
Higher secondary education (Havo/Vwo/Mbo)	25.8%	27.1%	39.4%
Tertiary education (Hbo/WO)	13.4%	13.1%	22.1%
Otherwise	1.4%	1.4%	0.1%
Total	100.0%	100.0%	100.0%

[a] The weighted figures constructed from the Agis HS 2001 are adjusted (direct standardization) to the 2001 Dutch national age and gender distribution of the sickness funds population (averaged over 12 months), as published by CVZ (2001). Age classified in the categories 15-24, 25-34, 35-44, 45-54, and 55-64.
[b] The CBS (2004) figures are based on the total Dutch population of 15-64 years of age, i.e. both sickness fund members (about two-thirds) and those privaletly insured.

population 20.2% have elementary education, 50.5% lower secondary, 24.9% higher secondary, and 4.2% tertiary (0.2% unknown). Thus, the educational level of the Agis survey sample is in between that of the Dutch national population and that of the Z&Z survey sample.

According to CBS (2004, section 2.2, Table 3), 50% of the Dutch households had disposable incomes of € 22,245 and above in 2000. From Table 3.5 disposible household income appears to be € 21,781 and above for 24.7% of the Agis sickness fund members in 2001. The main explanation for this difference is the fact that in order to become a sickness fund member in 2001, individual gross income

Table 3.5: Household income

Variable	Agis HS 2001	Agis HS 2001, standardized [a]	CBS 2000 [b]
Less than € 5,445 net a year	3.9%	4.6%	<50%
From € 5,445 up to € 10,891 net a year	19.4%	19.2%	
From € 10,891 up to € 16,336 net a year	30.8%	30.4%	
From € 16,336 up to € 21,781 net a year	21.2%	21.2%	
From € 21,781 up to € 27,227 net a year	11.4%	11.4%	>50%
From € 27,227 and more net a year	13.3%	13.2%	
Total	100.0%	100.0%	100.0%

[a] The weighted figures constructed from the Agis HS 2001 are adjusted (direct standardization) to the 2001 Dutch national age and gender distribution of the sickness funds population (averaged over 12 months), as published by CVZ (2001). Age classified in the categories 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85 and older.

[b] These CBS (2004, section 2.2 Table 3) figures are based on the 2000 total Dutch population, incl. the subset of sickness fund members (about two-thirds). Furthermore, the Agis HS 2001 records are sampled from a population of enrollees of 16 years and older.

has to be below € 29,813. As a consequence only two-thirds of the total Dutch population is a sickness fund member, the others have a private insurance.[58]

3.2.3 Selection of eligible cases for the analysis sample

In Table 3.6 gross response of 23,163 enrollees is reported. However, not all 23,163 records appear to be eligible for analysis. There are two reasons for exclusion of a record: responses appear not to be given by the Agis enrollee intended to be surveyed or the records turn out to be incomplete. In order to determine whether answers are given by the person intended to be surveyed, self-reported gender and date of birth are compared to the administrative records. For 487 persons these records did not match.

In order to determine whether a record is complete or not, the formal procedure that is devised with respect to conducting the CAHPS 3.0 Adult Commercial Questionnaire (CAHPS 2002) is applied. The first step in this procedure is to flag the so-called key questions. Key items include questions that all respondents should answer. In our case, the key questions are the SF-36 items, questions 12, 13, 14, 15 and 16 on sickness, questions 17 (excl. third item on most important reason for visit), 18, 19, 20, 21, 22 (only the yes/no item), 23 (only the yes/no item), and 24

58. A validity check is performed by multiplying the Agis percentages presented in the Table by (2/3). In that case the resulting percentage for sickness members with disposable household income below € 21,781 matches with the national percentage of 50%. However, this validity check is formally not correct as the two-thirds criterion applies to individual sickness fund members, whereas the percentages presented in the Table are calculated at the household level.

Table 3.6: Identification of eligible cases to be included into the analysis sample.

Sample category	Number of included records	Number of excluded records	Motivation for exclusion
Gross response	23,163		
		487	Responses not given by intended member [a]
		647	Answers given to less than half of the key questions [b]
Net response	22,029		
		2,723	SF-36 scores could not be calculated or imputed
		5	Missing record in WOVM 2001 database [c]
		82	Invalid record in WOVM 2001 database [d]
		589	Missing record in WOVM 2002 database [e]
		13	Invalid record in WOVM 2002 database [d]
Analysis sample	18,617		

[a] As indicated by date of birth and gender.

[b] The 137 key questions are: The SF-36 items, questions 12, 13, 14, 15 and 16 on sickness, questions 17 (excl. third item on most important reason for visit), 18, 19, 20, 21, 22 (only the yes/no item), 23 (only the yes/no item), and 24 on health care use, the OECD items of question 25, and all questions 26 through 37 on the background characteristics. Each question item is treated as a separate key question, except for questions 33 and 37.

[c] As indicated by year of birth.

[d] Based on the value of the error indicator 'recrd_ok' as determined by Vektis.

[e] As indicated by the date of birth, gender and ZIP-code (the first four numeric positions).

on health care use, the OECD items of question 25, and all questions 26 through 37 on the background characteristics. Each question item is treated as a separate key question, except for questions 33 and 37. If at least half of these 137 key question items (i.e. 69) are filled out, then the record is accepted as complete and therefore eligible for analysis. The 50 percent cut-off is a choice the CAHPS team has made to guarantee a uniform definition of a complete questionnaire. Excluding incomplete records as well as records containing responses from persons not intended to be surveyed leaves a net response of 22,029 records.

As in this study the SF-36 scale scores and the administrative data 2001 and 2002 are crucial, additional checks are applied to the records. It should be possible to calculate or impute the SF-36 scales for an enrollee, and 2001 and 2002 administrative records should be present and valid. In Chapter 4 a more in-depth analysis will be presented with respect to the calculations that were done in order to construct the SF-36 scales. It turns out that 2,723 records have to be dropped from analysis because of missing SF-36 scale scores. After validation of the administrative records, in the end 18,617 records remain for analysis.

3.3 External data sources

In this section the data from external data sources are described. Exclusively for this study, these data are kindly provided by the Dutch research and consultancy companies Aarts, De Jong, Willems, and Goudriaan Public Economics B.V. (APE) located in The Hague, and the Prismant Foundation (Prismant) located in Utrecht.

- The Dx Groups (DxGs) for the respondents are obtained from Prismant, based upon the inpatient diagnostic codes 2001 that are obtained from individual Dutch hospitals and pooled into the so-called LMR registration. In order to derive the Diagnostic Cost Groups (DCGs) at the individual level, the 1999 Dutch classification of DxGs into DCGs is applied. In the current study it turns out that there are no members associated with the thirteenth DCG, so we restrict ourselves to 12 DCGs.
- Indicators of health, medical supply, and consumption tendency regarding the year 2002 are obtained from APE (variable names between brackets):
  1. The number of sickness fund members in 2002 per ZIP-code, differentiated by age, gender, and eligibility are obtained from the Dutch Agency for Health Insurance (CvZ).
  2. Demographic data at the ZIP-code level are extracted from the CBS Statline website http://statline.cbs.nl. Dated at 1st of January, 2003. (Immigrants, sickness fund enrollees, single person households)
  3. Urbanization data at the ZIP-code level are obtained from the CBS via CvZ. Dated at 1st of January, 2004. (OAD)
  4. Socio-economic data at the ZIP-code level are obtained from the "Regional Income Research 2000" dataset of the CBS. (Low incomes)
  5. The distances from the four-digit ZIP-code of the sickness fund enrollees' residence to the four-digit ZIP-code of the health care providers, are based on the "Travelmanager centroids 2003" of ANDES VSP.[59] (distance to hospital, distance to general practitioner)
  6. Data per ZIP-code on the number of general practitioners for outpatient services are obtained from NIVEL. Dated at January 1, 2003.
  7. Standardized mortality rate (SMR) at the ZIP-code level is derived from the 1999 up to 2001 WOVM databases. The SMR is based on the information about the number of sickness fund members that were no longer enrolled at the end of the year because of death. A three years average is applied

59. Calculations are based on distances between the centroids of the four-digit ZIP-code areas. The "Travelmanager centroids 2003" contains the X- and Y-coordinates of all four-digit ZIP-codes in the Netherlands.

in this research. If the number of insured members is less than 100, the SMR of this ZIP-code is set equal to that of the municipality instead. For the years 1999, 2000, and 2001, weights of 0.2, 0.3, and 0.5 are applied. (SMR)

8. Data on hospitals and nursing homes are obtained from Prismant. Dated at January 1, 2002 and 2001, respectively. (Hospital beds, nursing home beds)

Date of extraction of these indicators is January 1, 2003 where possible. This date almost coincides with the date of extraction of December 31, 2002 that holds for the personal characteristics of the enrollees present in the WOVM 2002 database that is used in this study. The APE indicators are merged with the other datasets used in this study, where the four-digit ZIP code is the key variable.

The regional classification that is used in the Dutch REF model is constructed by APE. Since 2002 the regional variation in health care expenditures is no longer captured by a classification of four-digit ZIP codes into five classes according to the degree of urbanization (OAD), but by a four-digit ZIP code classification of differences between actual costs and REF predicted costs. These REF predicted costs are calculated after estimation of equation (2.1), where the set of risk adjusters $\{X_{j,t-1}, j=1,..,J\}$ consists of age, sex, insurance eligibility, PCGs and DCGs (i.e. exclusive of a regional variable).

The resulting differences are aggregated to the four digit ZIP code level and then regressed on health status and health care variables for which it is assumed that in the short term they can (almost) not be influenced by insurers' policies and therefore compensation is needed. A Ward (1963) clustering procedure is applied to these estimated differences afterwards, such that only ten clusters of four-digit ZIP codes remain. This clustering procedure takes account of mean costs per cluster (such that costs for ZIP codes that are close to each other in terms of costs are put in the same cluster) as well as variance within and between the ten ZIP code clusters (such that the variance between clusters relative to the variance within clusters is maximized).[60]

The resulting classification of the ten ZIP code clusters then determines the creation of ten regional 0/1 dummy variables that are included as risk adjusters in the Dutch REF model, in addition to age, sex, insurance eligibility, PCGs and DCGs. In order to arrive at the definite REF weights, equation (2.1) is re-estimated with all risk adjusters including the APE regional dummy variables.

60. The APE regional clustering that holds with respect to the 2004 Dutch REF model is based on claims data prior to 2002.

- The indicators of health, medical supply, and consumption tendency listed above are obtained separately from APE as well. In our study these indicators are merged with the survey and claims data on the basis of the four-position ZIP-code that can be found in the WOVM 2001 data set.

3.4 Conclusions

In this study, data are used that are derived from administrative sources, from the tailor-made Agis Health Survey 2001, and from external research institutes. The contents of these data sources are described in this chapter.

Section 3.1 describes the panel dataset containing the claims data for Agis sickness fund enrollees during the period 1999-2002. In this study, the versions of the administrative datasets are used, that were validated and applied in the context of the Dutch REF models over the years.

Section 3.2 describes the construction of the Agis Health Survey 2001. In Section 3.2.1 the considerations are given with respect to the questions to be included, given the choice of a postal questionnaire as the preferred mode to measure self-reported health. Gross response of 23,163 questionnaire forms resulted after having sent out 50,022 in total. From a response and nonresponse analysis it is concluded that for the purpose of our study no standardization for selective nonresponse is needed.

In Section 3.2.2 statistical representativeness is tested by comparing health care contacts, self-reported diseases, health status, and lifestyle of respondents to the Agis Health Survey 2001 with Dutch national figures. It is concluded that especially long-term diseases are more prevalent amongst Agis enrollees. However, this may be explained by the fact that in the Agis Health Survey 2001, enrollees younger than 16 are excluded. Furthermore, national figures include people that have a higher income than the sickness fund threshold of € 29,813. As compared to the national averages, the Agis survey sample includes relatively more non-immigrants, enrollees with lower education and lower disposable household income.

Section 3.2.3 presents the selection of data records that are eligible for the analysis sample in this study. Given the gross response of 23,163 records, a net response of 22,029 records remains after applying the restriction of validity and completeness of the survey records. In order to determine the completeness of a survey record, the formal procedure that has been devised with respect to conducting the CAHPS 3.0 Adult Commercial Questionnaire (CAHPS 2002) is applied. For the analysis in this study 18,617 records can be used because the SF-36

scale scores could be derived, and 2001 and 2002 administrative records are both available and valid.

In Section 3.3 the data obtained from external data sources are described. Appendix A3.6 summarizes the variables from the three types of data sources that are available for this study.

Appendix A3.1: Mode of data collection

In-person interviews are generally regarded as the survey mode of data collection to be preferred. They are traditionally considered to yield higher response rates, less nonresponse bias and better data quality. Nevertheless, mainly because of increasing costs, data are frequently collected by mail or telephone nowadays. In this appendix three studies are discussed in which different modes of data collection are compared. Based on these studies, it is decided that the mail survey is the mode of data collection to be preferred for the purposes of this study.

Van Campen et al. (1998) conclude that relatively few international studies compared the modes of administration with respect to health surveys. In total, four studies compared three modes of data collection, and nine studies compared only two modes of data collection. The modes were compared in terms of the total costs, the survey response rates and data quality in terms of completeness of the data. Most studies dealt with health status.

Total survey costs per case were highest for the in-person interviews (approximately US$ 55). The telephone survey costs approximately US$ 29 per case, while the total survey costs of the mail survey are lowest (approximately US$ 10).

The survey response rates were highest for the mail survey (85%). The response rates for the telephone (53%) and in-person (61%) interviews were much lower. The response rates of the elderly (64%) were lowest, particularly in the telephone interview mode (25%).

The least missing values per item was best in the telephone interviews (all respondents scored less than 10% missing values, calculated over the total number of 55 items), the in-person mode performed second-best but not significantly different (95% of the respondents scored less than 10% missing values). With the mail interview 81% of the respondents scored less than 10% missing values, significantly worse than both other modes. The consistently higher proportion of missing values in the responses of the elderly patients confirms the general observation of the difficulties of interviewing elderly people.

With respect to data quality, mixed results are reported for tests of reliability and validity of scores on perceived health status and health services utilization across survey methods. The answers to the health dimensions in Van Campen et al. (1998) did not differ significantly, except for the physical functioning of the mail responders that appeared significantly lower.

In Table A3.1 we give a qualitative summary of the main conclusions.

Table A3.1: A qualitative evaluation of three modes of administration, based on Van Campen et al. (1998)

	Mail survey	Telephone interview	In-person interview
Data Collection Costs	+	±	-
Response Rates	±	±	+
Partial nonresponse	±	+	+
Data Quality	+	+	+

An interesting addition to this inquiry of Van Campen et al. (1998) is the study by Van Sonsbeek and Stronkhorst (1983), who compared in-person interviews, written interviews and mixed in-person and written interviews.[61] It should be noted that the mail survey mode described by Van Campen et al. (1998) differs from the written interview because the interviewer delivers the survey personally at the home address and collects general household characteristics before handing over the questionnaire form.

Van Sonsbeek and Stronkhorst (1983) did not discuss the costs associated woth the three modes of data collection. They assume that the data collection costs for the written interviews are not much lower than those for the in-person and mixed interviews, because these include the costs of the interviewer that pays a second home visit in order to collect the questionnaire form. Usually, in-person interviews are most costly.

Response rates are highest for the in-person interviews, about four percentage points higher than the other two modes of data collection. In big cities, nonresponse is generally above average for all three modes of data collection. From a nonresponse analysis it appears that the reasons for nonresponse are not significantly different between the three modes of data collection.

The written interviews show highest partial nonresponse rates with respect to almost all survey questions, the mixed interviews show smallest partial nonresponse. It is assumed that an interviewer can play a positive role with respect to the reduction of partial nonresponse, i.e. missing value rates and routing errors. The rationale for this assumption is that the interviewer can explain concepts and motivations, which is especially important for the elderly, the lower social-economic categories, immigrants, etc. However, Van Sonsbeek and Stronkhorst (1983) argue that interviewers can be a source of errors as well, amongst others because of making suggestions to answers, wrong interpretation of concepts,

61. With the mixed interviews, all persons present at home are interviewed personally. For the household members that are not present at that moment, it is requested that they fill out a questionnaire form themselves that will be collected after two weeks.

amount of experience. There are different opinions on to the amount of errors that is introduced this way.

It appears that some questions have high partial nonresponse for all three modes of data collection, especially insurance policy related questions, questions on the length of stay in a hospital, income and use of maternity care. Answers to questions on attitudes towards and perceptions of health are given more often. It should be noted that in our study, the choice of the mode of collection is largely determined by the measurement issues with respect to the subjective information. Information on medical consumption is largely derived from the claims data.

The answers to the question on perceived general health status, differ significantly between the written interview on the one hand, and the in-person and mixed interviews on the other. These differences appear especially with respect to the categories "very good" and "good" health: the answers given during the in-person interviews are biased to the "very good" health category, perhaps because of politeness with respect to the interviewer.

Although combining the categories "very good" and "good" health eliminates the differences between the modes of data collection, it is not a recommended policy because significant differences in related health variables are reported between these categories (e.g. long-term disorders and primary care contacts). Furthermore, for each separate category, there do not appear to be significant differences between the three modes of data collection in these related variables.

There appears to be no difference in data quality as the self-reported number of GP visits, medical specialist visits, physiotherapist visits, dentist visits and hospital admissions and drugs utilization are comparable between the modes of data collection. However, in some cases deviations from external data sources (e.g. registrations) seem to exist.

In Table A3.2 the main conclusions are summarized.

Table A3.2: A qualitative evaluation of three modes of administration, based on Van Sonsbeek and Stronkhorst (1983)

	Written interview	**Mixed in-person and written interview**	**In-person interview**
Data Collection Costs	±	-	-
Response Rates	±	±	+
Partial nonresponse	±	±	±
Data Quality	+	-	-

The RAND Health Insurance Experiment relied heavily on self-administered forms because they are far less expensive than personal interviews and because it was assumed that respondents might be willing to put down sensitive health status

information on a self-administered form that they might not be willing to tell an interviewer (telephone interviews were only used as a last resort). From Newhouse (1993) it appears that an overwhelmingly 70% of the participants preferred self-administered questionnaires to personal interviews (12% preferred personal interviews, 19% had no preference). Given the additional expense of personal interviews and the demonstrated quality of the data obtained from self-administered forms, Newhouse (1993) recommends that analogous future efforts should use self-administered forms.[62]

In conclusion, we prefer the self-administered mail survey mode in order to collect information on health status.[63] This mode of administration scores best on data collection costs and data quality of questions on attitudes towards and perceptions of health. The response rates and partial nonresponse are slightly better in other modes of data collection, but not in all studies.

For the present study, the survey is sent to Agis enrollees of 16 years and older. The Lamers (1997) study makes use of the results of a mail survey conducted among enrollees of the Dutch sickness fund Zorg & Zekerheid (Z&Z), where questionnaires were also sent to parents and caretakers to fill out the mail survey for children up to 16 years. In our research such proxy-surveys have not been conducted, in order to stay within the budgetary limits that hold for this study. There are two arguments in defense of this decision. First of all, with interviews, Cannell (1997) observes that medical consumption is usually underreported, especially with so-called proxy-interviews. Furthermore, subjective information like attitudes and perceptions cannot be gathered reliably from proxy-interviews. Therefore, Statistics Netherlands (CBS) no longer conducts proxy-interviews since 1997. The second reason is that Dutch insurers do not charge premiums for people under 18 years of age, because a state subsidy holds for basic coverage of children instead. Therefore, insurers do not set premiums according to their younger enrollees' risk. As financial access to coverage for children is safeguarded in this way, the impact of risk equalization models with respect to adults is of main concern in the applications of the test procedure proposed in this study.

---

62. It should be noted that the participants had to fill out biweekly as well as annual questionnaires. As a reminder to their mail questionnaires, 48% of the respondents did state no preference to either a letter or a phone call, a reminder by phone call was preferred by 30% of the respondents, and 22% preferred a letter.

63. Note that for the purposes of our study we need information on health status for enrollees with and without health care utilization in 2001. Therefore sampling health status data from medical files was an option.

Appendix A3.2: Predictive power of Z&Z survey variables

May 2000 a regression of 1994 health care expenses of the Dutch sickness fund Zorg & Zekerheid (Z&Z) was estimated, with available health survey variables taken as explanatory variables.[64] The reason for conducting this regression was to determine which survey questions have significant explanatory power and thus should be included into the Agis Health Survey 2001. The models estimated are a 1994 model with total hospital expenses and a 1994 model with variable hospital expenses only, where variable hospital expenses are defined according to the so-called "splitsings" model as applied in the Netherlands (with 1996 tariffs). For those questions that did not apply to children, the corresponding observations were automatically removed in the stepwise regression procedure.

The personal judgment on present general health significantly explains part of the variation in health care expenses. General health five years ago and the question on how many days one had to stay in bed the past 6 months have only significant effects in the so-called "splitsings" model.

The VOEG and ABS score are called old-fashioned and scarcely used by Statistics Netherlands (CBS), see Van den Berg and Van der Wulp (2003). Furthermore, from the regression it appears that the physical dimension of the VOEG (which is largely measured by the VOEG instead of the psychological dimension) does not contribute significantly to the explanation of the variation in health care expenses, just as the psychological ABS questions are redundant. Although three items which represent the psychological dimension of the VOEG score do have a significant contribution to the explanation of the variation in health care expenses, it is no option to include only 3 out of 25 VOEG items in the survey.

The Quetelet-index does not have a significant contribution to the explanation of health care expenses variation. Thus, length and weight do not have to be included into the Agis Health Survey 2001.

Total number of treated disorders do appear to significantly explain the variation in health care expenses, as well as the OECD items, the ADL items and the propensity for consumption. It should be noted that the propensity for consumption correlates significantly with use of health care facilities.

64. The sample used for this analysis is described in detail in Lamers (1997).

Appendix A3.3: Data collection process

In the first two subsections, the choice for the postal survey mode is discussed and the selection of the survey questions to be included into the Agis Health Survey 2001. In this subsection, the data collection process is presented. In order to setup this process, attention is paid to the statistical theory on sample designs, the Dutch legal preconditions are studied in order to be allowed to conduct a health survey amongst sickness fund members, and a choice is made for the survey vendor to send out the survey and to record the survey answers. The data collection process will be described, both for the pilot and the main surveys that were conducted.

Pilot survey

A pilot survey amongst 1000 Agis beneficiaries is conducted in the period between April 25 and June 26, 2001. The purpose of this pilot was to test all stages of the survey procedure in order to determine the strong and weak points of the procedure that was set-up. With those results it is possible to make procedure adjustments in order for the main survey to be conducted adequately. In addition, with this pilot it was tested whether a financial incentive would increase response rates.

From the so-called WOVM 1997 and 1998 databases with more than 1.6 million Agis members each, first only those members are selected that appeared in both databases.[65] In addition, the database is restricted to those that are between 16 and 90 years of age in 2001. A stratified sample of 2000 Agis beneficiaries is drawn, in order to be sure that there would remain at least 1000 members after the final match with the membership administration the day before posting the questionnaire forms (i.e. at April 25, 2001).

The survey sample is stratified proportional to one-year predicted health care expenses in 1998. In order to oversample beneficiaries with chronic diseases and conditions, one-year predicted expenses of 1998 are used as a sampling weight. For that purpose, a linear regression of total expenses 1998 on the following 1997 predictors is fitted by ordinary least squares: age, gender, eligibility, region, 22 PCGs (not rank-ordered), hospitalization, and dummies that indicate whether someone belongs to the top 5% group of members with highest expenses for medical devices or paramedic services, or has had any rehabilitation expenses.[66]

65. See footnote 17 for an explanation of the abbreviation WOVM.

66. From the evaluation of the pilot phase, it turned out that the PCG "hypertension (low)" accidentally has not been included in the aforementioned regression.

These variables are based on health care expenses data 1997 and 1998, and 1997 pharmaceutical recipes containing information on ATC codes and DDD. Given the estimated regression parameters, a prediction of 1998 expenses can be made which functions as the sampling weight in drawing the survey sample. Length of membership has not been accounted for in this case.

Hospitalization is a dummy that indicates whether someone was hospitalized. Clinical morbidity could be captured more precisely by replacing this dummy by indicators of medical specialisms that were contacted. However, this is not possible as the necessary Anova 1997 claims data appear not to be complete.

In addition, for roughly half of the Anova beneficiaries the pharmaceutical claims data are lacking, so the PCGs for these members equal zero by definition. Therefore, their sampling weight when drawing the survey sample will be lower than that of ZAO and Anoz members with comparable health status, and as a consequence there will be relatively fewer Anova members with chronic conditions in the pilot sample.[67]

Note that although there may exist a positive correlation between the sampling weight and the SF-36 health status scores, this will not invalidate the research analysis as they are measured at different moments in time. Furthermore, these weights are mainly indicators of chronic conditions which are supposed to be time invariant.

At April 18, 2001 a database with postal addresses of 1,548 Agis members is sent to the survey vendor by e-mail. The first 1,030 records are selected, after sorting the database by the stratification weight. As the stratification weight is larger for Agis members for which predicted expenses are larger, a subgroup of members with the largest predicted expenses has not been included in the survey sample. As a consequence, those members who are admitted to a hospital, or those who suffer from Parkinson, diabetes type I, cancer, cystic fibrosis, (end-stage) renal disease, HIV/aids, or have had a transplantation in 1997, are not present in this sample. As such, this procedure has led to the non-representativeness of the pilot sample. In the main survey, no sorting of any kind is applied to the database with postal addresses in order to avoid this type of error.

Next the records were split in two separate mailings, one with a PleisterClowns contribution response incentive, the other without. Splitting is done one record after the other. Note that because of the sorting issue the response incentives

67. In the main survey, the complete pharmaceutical claims database of Anova will be used. Note that the conclusion does not hold with respect to hospitalization, as this variable is based on the complete WOVM 1997 database.

may be attributed non-randomly to the Agis members in the pilot sample. It is not possible however, to determine to what extent this will diffuse any results.

Before the final posting, another 30 records were deleted in order to come to a total amount of exactly 1000 pilot records. The survey vendor, however, inappropriately deleted all 30 records from the database with Agis members that are mailed including the ClinicClowns response incentive. As a consequence, 485 members were mailed including the response incentive and 515 excluding the response incentive.

Upon first delivery of the database with survey responses to the Erasmus University, the survey vendor appeared to have mailed the digitally encoded respondents only. Furthermore, the survey vendor also did not send a coding book containing a description of the variable codes. Moreover, matching the survey records with the administrative records had to be done by ZIP-code and house number, as the key was lacking.[68] Two weeks after delivery, a complete database including non-respondents, and the connection key was delivered.

It can be concluded that the pilot survey was indispensable as a means to determine and remove the weak points of the sampling procedure, both with respect to Agis and the survey vendor. It is recommended that sampling procedures be certified in case of structural use of survey vendors for these kind of activities. Inappropriate procedures may lead to invalid survey data and therefore probably to invalid research results.

Main Survey

As was also the case with respect to the pilot survey, with respect to the main survey it was our intention to oversample beneficiaries with relatively high predicted 1998 expenses.[69] This stratification procedure was not applicable to all Agis members, because it was not possible to calculate predicted 1998 expenses for the Agis members that had contracted Agis after 1998. With respect to the main survey the population therefore is split into two strata, the first consisting of members already being insured in 1997 and 1998 and were still insured at the start of September 2001 (1,129,463), and the second stratum consisting of those members that became insured after 1998 and were still insured at the start of

68. The survey vendor has constructed a database containing the connection key between the administrative identification numbers and the questionnaire forms numbers.

69. There were no data available more recently than 1998 in order to predict health care expenses.

September 2001 (126,514). Stratified members are at least 16 and at most 90 years of age at the time, and not institutionalized.[70]

In Table A3.3 the estimated coefficients are presented for the regression of Agis enrollees that have already been insured in 1997 and 1998 and are still insured at the start of September 2001. The sampling weights in the first stratum are equal to expected expenditures, given these estimated coefficients. Equal sampling weights are applied in the second stratum, consisting of those members that became insured only after 1998 and still insured at the start of September 2001.

From the first stratum 47,000 mail addresses are sampled following the same stratification procedure that is applied in the pilot survey[71]; from the second stratum 3,100 members are uniform randomly sampled.[72] In the end 46.979, and 3.043 sampled units remained, as some records are deleted afterwards because of the following restrictions:

- At most two members within one household are surveyed;
- At October 3, 2001 the selected members to be surveyed are checked for their membership being still valid;
- No member already included in the pilot should also be included into the main survey.

On October 3, 2001 a database with 50,022 sampled postal addresses was sent to the survey vendor. From October 11, 2001 to January 18, 2002 questionnaire forms were mailed to 50,022 Agis members following the Dillman (1978) mailing procedure. October 11, 2001 a questionnaire form was sent out to these 50,022 Agis members, with a cover letter explaining the purpose of the health survey. The cover letter contained both the logo of Agis Health Insurance and Erasmus University Rotterdam, and was signed by representative persons of both organizations. At the end of the second week, i.e. on October 24th, a post card was sent as

---

70. Agis members in old people's homes and institutions for daily stay were excluded from the sampling procedure for the purpose of the main survey. The exclusion was based on ZIP-codes and home numbers of these so-called AWBZ institutions. However, in the pilot, records with respect to Agis members living in nursing or old people's homes were excluded manually *after* the sample of enrollees for the pilot survey was drawn. Although a database containing ZIP-codes and house numbers of these institutions was available for the pilot, the format was not yet appropriate for digital processing at the time. The list of institutionalized enrollees was made ready for digital processing after the pilot survey.

71. However, in contrast to the pilot survey, the linear regression is weighted by length of membership this time.

72. The results with respect to the second stratum make it possible to test whether results for these new members are comparable to the results of the so-called "non-movers" of the first stratum.

Table A3.3: Estimated coefficients of 1997 explanatory variables in a linear regression with total 1998 health care expenses in Euro as dependent variable, weighted by length of membership.

Explanatory variable	Parameter Estimate
Intercept	279 *
Male	0
Female	123 *
15-19	34
20-24	58 *
25-29	118 *
30-34	154 *
35-39	0
40-44	0
45-49	175
50-54	218
55-59	268
60-64	332 *
65-69	694 *
70-74	853 *
75-79	958 *
80-84	783 *
85-89	747 *
Disabled 15-34	905 *
Disabled 35-44	893 *
Disabled 45-54	579 *
Disabled 55-64	443 *
(Self-)Employed 15-34	0
(Self-)Employed 35-44	107 *
(Self-)Employed 45-54	-53
(Self-)Employed 55-64	-51
Social welfare 15-34	205 *
Social welfare 35-44	349 *
Social welfare 45-54	195
Social welfare 55-64	178
Unemployed 15-34	127 *
Unemployed 35-44	241 *
Unemployed 45-54	51
Unemployed 55-64	2
Retired	-35

Explanatory variable	Parameter Estimate
OAD1 [a]	0
OAD2 [a]	-18 *
OAD3 [a]	-40 *
OAD4 [a]	-49 *
OAD5 [a]	-68 *
Hypertension (low)	627 *
Glaucoma	237 *
Gout	579 *
Thyroid disorders	230 *
Tuberculosis	1508 *
Hypertension (high)	170 *
Depression	418 *
Diabetes Type II	433 *
Hyperlipidemia	556 *
Respiratory illness, asthma	1100 *
Epilepsy	834 *
Rheumatologic conditions	934 *
Cardiac disease / ASCVD / CHF	1228 *
Crohn's and ulcerative colitis	1217 *
Acid peptic disease	1317 *
Parkinson's disease	1845 *
Diabetes Type I	1537 *
Cancer	3173 *
Transplantations	5027 *
HIV/Aids	12847 *
Cystic fibrosis	3842 *
Renal disease (including ESRD)	29895 *
Hospitalization [b]	1621 *
Medical devices (top 5% expenses)	1198 *
Paramedics (top 5% expenses)	708 *
Rehabilitation	4550 *
R^2_{ADJ}	11.09%

[a] OAD is an abbreviation for the address density of the surrounding area, which is a measure of urbanization (see Den Dulk, Van der Stadt, and Vliegen, 1992). Density for enrollees living in the OAD1 cluster is highest, and lowest for those living in cluster OAD5.

[b] Dummy variable that equals one for those enrollees with a positive number of hospitalizations and at least one recorded hospitalization day, zero otherwise.

a reminder to everyone mailed in the first wave. Respondents were thanked for their cooperation, non-respondents were kindly asked to send back their survey this time. Three weeks after this, November 15th, all non-respondents received a new questionnaire form accompanied by a short cover letter. A final, fourth mailing consisting of a one-page letter was sent two weeks after this at November 29th to all Agis members who had not responded until then.[73]

The survey vendor has digitally encoded the returned questionnaire forms and has come up with a database containing 24,129 records with survey answers. Some of those records appeared to be identical, some completely blank, and in some cases two records had to be merged into one record. In other cases, enrollees appear to have filled out the questionnaire twice but answers are not identical with respect to every survey question. In this situation, the record is selected that was received by post first. If received on the same day, the record that was scanned last are selected, unless filled out much worse than the first one. In the end, 23,163 unique records remain for analysis. In Table A3.4 the number of included and excluded records is listed, given the selection criteria applied to the 50,022 records associated with enrollees that received a questionnaire form.

Table A3.4: Gross response to the questionnaire.

Sample category	Number of included records	Number of excluded records	Motivation for exclusion
Mailed	50,022		
		25,893	Questionnaire form not returned [a]
Returned	24,129		
		2	Key number missing in record
		814	Completely blank record
		122	Key number appears in multiple records
		28	Multiple records that match exactly
Gross response	23,163		

[a] Inclusive of 42 members not corresponding to any record in the WOVM 2001 database.

To summarize, in this appendix the data collection process for the Agis Health Survey 2001 is described in detail. The Dillman (1978) four step mailing procedure is applied to collect the survey data, also with respect to the pilot survey amongst 1000 enrollees that is held in Spring 2001. The pilot survey is held in order to test the mailing procedure itself, as well as to test whether a financial incentive

73. Because of operational restrictions this fourth mailing was split up in three separate mailings, each sent out one week after the other.

would increase response rates. It is concluded that a pilot survey is indispensable as a means to determine and remove the weak spots of the sampling and mailing procedure, both with respect to Agis and the survey vendor. It is recommended that sampling and mailing procedures be certified in case of structural use of survey vendors for these kind of activities. Inappropriate procedures may lead to invalid survey data and therefore probably to invalid research results.

Appendix A3.4: Description of Agis Health Survey 2001 items

Table A3.5: Description of survey items presented in Tables 3.2 and 3.3, inclusive corresponding survey question number in Agis Health Survey 2001 [a]

Variable	Description	Survey question number
% Enrollees with GP contact	The percentage of sickness fund members having a contact per year	17a
% Enrollees with specialist contact	The percentage of sickness fund members having a contact per year, exclusive of the contacts during hospitalizations	18a
% Enrollees with hospitalization	The percentage of sickness fund members hospitalized per year	18b, item a
% Enrollees with paramedic contact	The percentage of sickness fund members having a paramedic (physiotherapy, César and Mensendieck) contact per year	19, items c and f
% Enrollees with RIAGG contact	The percentage of enrollees that *ever* had a contact with a Regional Institute for Ambulantory Mental Health Care (RIAGG). The percentage from the POLS survey reflects those that had a contact in *a year*.	20a
% Enrollees with alternative care contact	The percentage of sickness fund members having a contact per year with an alternative care practitioner (incl. Gps). Contact is defined as having had contact with any of the listed alternative care practitioner.	21, items a-g
Pharmaceutical drugs		
% enrollees with prescribed drugs	The percentage of sickness fund members taking prescribed drugs during the 14 days previous to the survey date	24a
% enrollees with non-prescribed drugs	The percentage of sickness fund members taking non-prescribed drugs during the 14 days previous to the survey date. Anticonceptive drugs and drugs used during hospitalizations are not taken into account.	24b
Long-term diseases		
Diabetes mellitus (Type I and II)	Including those still suffering or under treatment.	14a

Variable	Description	Survey question number
Stroke, brain haemorrhage/infarction	Including those still suffering or under treatment.	14b
Myocardial infarction	Including those still suffering or under treatment.	14c
Other serious heart disease	Including those still suffering or under treatment.	14d
Some type of (malignant) cancer	Including those still suffering or under treatment.	14e
Migraine or serious headache regularly	Including those still suffering or under treatment.	16a
Hypertension	Including those still suffering or under treatment.	16b
Vascular constriction (stomach, legs)	Including those still suffering or under treatment.	16c
Asthma, COPD	Including those still suffering or under treatment.	16d
Psoriasis	Including those still suffering or under treatment.	16e
Chronic dermatitis	Including those still suffering or under treatment.	16f
Dizziness when falling down	Including those still suffering or under treatment.	16g
Intestinal obstructions (> 3 months)	Including those still suffering or under treatment.	16h
Urinary incontinence	Including those still suffering or under treatment.	16i
Serious/persistent back problem	Including those still suffering or under treatment.	16j
Osteoarthritis (hip/ knees)	Including those still suffering or under treatment.	16k
Chronic joint inflammation	Including those still suffering or under treatment.	16l
Other serious/persistent injury (neck, shoulder)	Including those still suffering or under treatment.	16m
Other serious/persistent injury (elbow, wrist, hand)	Including those still suffering or under treatment.	16n
Other prolonged disease/ disorder	Including those still suffering or under treatment.	16o
Generic health status measure		
Physical Component Scale score	Zero scale score is worst health state possible, one hundred is optimal physical health.	
Mental Component Scale score	Zero scale score is worst health state possible, one hundred is optimal mental health.	
Psychological distress		
Fearful or afraid (for 2 months)	Including those still suffering or under treatment.	15a
Downhearted or blue (for 2 months)	Including those still suffering or under treatment.	15b

Variable	Description	Survey question number
Functional limitations		
% enrollees with one or more OECD limitations	The OECD indicator scores "yes" if at least one out of seven items is answered with "cannot" or "with much difficulty".	25, excl. item e
% enrollees with OECD auditive impairment	The OECD indicator of auditive impairments scores "yes" if at least one out of two items is answered with "cannot" or "with much difficulty".	25, items a and b
% enrollees with OECD visual impairment	The OECD indicator of visual impairments scores "yes" if at least one out of two items is answered with "cannot" or "with much difficulty".	25, items c and d
% enrollees with OECD mobility impairment	The OECD indicator of mobility impairments scores "yes" if at least one out of three items is answered with "cannot" or "with much difficulty".	25, items f, g, and h
Lifestyle		
Height	In centimeters (1.00 inch = 2.54 cm)	32
Weight	In kilograms	31

[a] Bedridden because of illness or injury (survey question 12) has not been presented in these tables, as the references periods in the Agis survey and the CBS POLS survey are incomparable.

Appendix A3.5: Response and nonresponse analysis

In this subsection response and nonresponse with respect to the Agis Health Survey 2001 is analyzed, both for the pilot and the main survey. A distinction is made between early and late respondents, early respondents being those that returned their questionnaire forms after having received the first or second mailing, late respondents after having received the third or fourth mailing.[74] This distinction is made in order to check whether the four-step mailing procedure recommended by Dillman (1978) was useful.

The pilot survey was conducted from April 25, 2001 up to and including June 26, 2001 amongst 1000 Agis members. Table A3.6 shows that 45% of the mailed

74. The variable Z_extra in the electronic survey database is used in order to classify respondents as early responders (i.e. wave 1 and 2) and late responders (i.e. wave 3 and 4). Z_extra contains the day number in the year 2001 on which the returned questionnaire form was delivered by the PTT to the survey vendor. The dates on which the mailings were delivered to the Dutch PTT Post Office by the survey vendor are 10/11 (wave 1), 10/24 (wave 2), 11/15 (wave 3) and 11/29 (wave 4). These dates correspond to Z_extra values of 284, 297, 319 and 333, respectively. Assuming that a questionnaire form can be filled out and returned within two days in theory, everybody for which the questionnaire form was received before day 323 of the year 2001 (Z_extra < 323) is classified as early responder ("Wave 1&2").

Agis members have returned the questionnaire form. Furthermore, men appear to respond less often than women.

Table A3.6: Total response, and early and late response by gender for pilot survey (unweighted) [a]

		Response percentages		
Gender	**Number of mailed forms**	**Early respondents**	**Late respondents**	**All respondents**
Male	399	25%	13%	38%
Female	601	32%	18%	49%
Total	1,000	29%	16%	45%

[a] From chi-square tests it appeared that response percentages differ statistically significantly between men and women for all subgroups of enrollees presented here (95% level of confidence).

In this study, the Dillman (1978) mailing procedure is followed. In order to determine whether the four-step procedure has paid off, total response to the first two mailings (29%) should be compared to that of the latter two (16%). However, the response of the non-responding members should be corrected for the number of enrollees already having responded after the first and second mailing, i.e. the response percentage amongst non-respondents to the first two mailings equals 16%/(1-29%)=23%. Comparing this figure to the 29% response rate with respect to the first two mailings, it may be concluded that responsiveness with respect to the third and fourth mailing is still quite substantial. Therefore, it may be concluded that the third and fourth mailing do have added value in terms of questionnaire forms returned.

From Table A3.7 it appears that the CliniClown incentive to respond did not improve response rates in the pilot survey. Therefore it is decided to conduct the main survey without the incentive.

Table A3.7: Total response, and early and late response by CliniClown response incentive for pilot survey (unweighted) [a]

		Response percentages		
Mailing type	**Number of mailed forms**	**Early respondents**	**Late respondents**	**All respondents**
Without incentive	515	30%	17%	47%
With incentive	485	27%	15%	42%
Total	1000	29%	16%	45%

[a] From chi-square tests it appeared that response percentages do not differ statistically significantly between men and women for all subgroups of enrollees presented here (95% level of confidence).

The questionnaire form itself has only slightly been adjusted based on the pilot experiences. The positioning of the questions on land of birth was altered, the question on marital status was rephrased, and based on the results with respect to the income distribution it was decided to insert an additional top income class in the main questionnaire. All the other questions have remained unaltered.

At Januay 18th, 2002, for the last time a questionnaire form was encoded by the survey vendor. From Table A3.8 it appears that 46% of the surveyed members have responded, i.e. gross response equals 23,163 out of 50,022. On average male members are less responsive than female members: 42% resp. 49%, analoguous to the outcomes in the pilot survey.

Table A3.8: Total response, early and late gross response by gender for main survey (unweighted) [a]

		Response percentages		
Gender	**Number of mailed forms**	**Early respondents**	**Late respondents**	**All respondents**
Male	19,632	27%	15%	42%
Female	30,390	33%	16%	49%
Total	50,022	31%	16%	46%

[a] From chi-square tests it appeared that response percentages differ statistically significantly between men and women for all subgroups of enrollees presented here (95% level of confidence).

In order to determine whether the four-step procedure has paid off in the main survey, total response to the first two mailings (31%) should be compared to that of the latter two (16%). Response rates of those enrollees that did not respond to the first and second mailing equals 16%/(1-31%)=23%. Comparing this figure to the 31% response to the first and second mailings, it may be concluded that also here the added value of the third and fourth mailing is substantial, rather similar to that for the pilot survey.

In this study, research results will be based on the survey answers of the respondents only. As selective nonresponse may have an impact on the statistics based on the survey outcomes, it is recommended to conduct a nonresponse analysis. A logit model is estimated in order to determine to what extent selective nonresponse exists in our study. This nonresponse analysis is applied to all surveyed Agis members.[75]

75. The logit model is based on 49,756 cases. From the total of 50,022 records, 266 records were excluded beforehand because their corresponding administrative characteristics were either not present at all in the WOVM 2001 database or entitled not valid by Vektis. There appear to be 109 respondents amongst these deleted cases. The null hypothesis of independence between nonresponse

In Table A3.9 odds ratio's and corresponding 95% confidence intervals are presented derived from a logit model that is estimated to determine the association between personal characteristics and tendency to respond to the survey. The personal characteristics comprise of those included in the conventional REF models as presented in Chapter Five. If the odds ratio of some personal characteristic is statistically significantly different from one, then the odds are against responsiveness for those enrollees as compared to those belonging to the reference group. For example, it appears that enrollees between 35 and 84 years of age have a higher tendency to respond than those between 25 and 34 years of age. On the other hand, enrollees of 85 and above tend to respond relatively less often. Furthermore, amongst men younger than 75 relatively less respondents can be found than amongst women of the same ages, for men of 75 and above the reverse holds.

Amongst those being disabled, on social welfare, unemployed and self-employed less respondents can be found than amongst employed enrollees. It should be noted that for each subgroup in Table A3.9 the odds ratio's reported are already corrected for the effects of other subgroups included in the regression. However, interaction effects between e.g. gender and employment status were included in this logit model. Therefore, no conclusion can be drawn about the odds to respond amongst the group of employed men.

Other personal characteristics included are the region where an enrollee lives and 25 medical conditions. Enrollees belonging to the first regional cluster tend to respond less frequently than those belonging to the second regional cluster, those belonging to the regional clusters five up to ten tend to respond more. For five out of twelve diagnoses derived from the pharmaceutical drugs claims, the odds were in favor of a response to the survey: asthma/COPD, cardiac disease, rheumatism, transplantation, and neuromuscular disorders). For three out of thirteen diagnostic costs groups, the odds are against responsiveness (clusters 5, 10, and 11).[76]

As in many studies the variable under study is available with respect to respondents only, selective nonresponse is usually analyzed with respect to all kinds of background characteristics instead of the variable under study. In this study, however, the focus is on 2001 health care expenses and these are available both for respondents and nonrespondents from the Agis administration. Therefore, here

---

and ineligibility of cases could not be rejected at the 5% confidence interval (Person chi-square test statistic equals 3.05, p=0.08).

76. In the logit model it is possible that enrollees belong to more than one subgroup of diagnoses.

Table A3.9: Odds ratio's and 95% confidence intervals for the REF risk adjusters as explanatory variables for response versus nonresponse to the main survey.

Explanatory variable	Odds ratio	95% confidence interval
15-24	1.03	0.92 - 1.16
25-34	1.00	---
35-44	1.30	1.20 - 1.42
45-54	1.56	1.43 - 1.70
55-64	1.95	1.78 - 2.14
65-74	1.87	1.50 - 2.33
75-84	1.37	1.10 - 1.71
>=85	0.74	0.58 - 0.94
M 15-24	0.60	0.52 - 0.70
M 25-34	0.60	0.54 - 0.66
M 35-44	0.53	0.49 - 0.59
M 45-54	0.66	0.60 - 0.73
M 55-64	0.72	0.65 - 0.79
M 65-74	0.90	0.83 - 0.98
M 75-84	1.12	1.01 - 1.24
M >=85	1.37	1.09 - 1.72
Disabled	0.69	0.65 - 0.74
Employed	1.00	---
Social welfare	0.53	0.49 - 0.58
Unemployed	0.72	0.66 - 0.80
Retired	0.88	0.72 - 1.08
Self-employed	0.78	0.68 - 0.90
APE Region 1	0.93	0.86 - 0.99
APE Region 2	1.00	---
APE Region 3	1.05	0.98 - 1.12
APE Region 4	1.07	0.99 - 1.14
APE Region 5	1.10	1.03 - 1.16
APE Region 6	1.16	1.08 - 1.24
APE Region 7	1.26	1.19 - 1.34
APE Region 8	1.24	1.10 - 1.40
APE Region 9	1.25	1.11 - 1.40
APE Region 10	1.25	1.15 - 1.37
Asthma/COPD	1.07	1.00 - 1.14
Epilepsy	0.90	0.76 - 1.06
Crohn/Colitis Ulcerosa	1.27	0.92 - 1.76
Cardiac disease	1.08	1.01 - 1.16
Rheumatism	1.59	1.19 - 2.13

Explanatory variable	Odds ratio	95% confidence interval
Parkinson	0.98	0.72 - 1.33
Diabetes (Type I)	1.00	0.91 - 1.10
Transplantation	1.61	1.15 - 2.27
Cystic fibrosis	1.42	0.76 - 2.65
Neuromuscular disorder	1.80	1.07 - 3.02
HIV/Aids	1.08	0.73 - 1.60
Renal disease/ESRD	1.00	0.56 - 1.76
DCG1	0.91	0.75 - 1.10
DCG2	1.15	0.95 - 1.39
DCG3	0.85	0.70 - 1.03
DCG4	0.89	0.73 - 1.08
DCG5	0.66	0.53 - 0.83
DCG6	1.03	0.75 - 1.42
DCG7	0.96	0.75 - 1.23
DCG8	0.99	0.74 - 1.32
DCG9	1.08	0.61 - 1.91
DCG10	0.47	0.31 - 0.71
DCG11	0.49	0.31 - 0.78
DCG12	1.04	0.69 - 1.58
DCG13	0.80	0.53 - 1.20
Pseudo-R^2	3.69%	

it is also possible to analyze selective nonresponse with respect to the variable under study.

In Table A3.10 average health care expenditures of respondents are compared to those of non-respondents for specific categories of health care expenditures (except GP). After indirect standardization (age, gender, eligibility, region, pharmaceutical and clinical diagnoses), it appears that total expenses do not differ statistically significantly between respondents and non-respondents, the same conclusions holds for those being hospitalized. On the other hand, average expenses appear significantly lower amongst non-respondents receiving pharmaceutical drugs, paramedic care, dental care or medical devices.

When nonresponse is selective, research results may need standardization in order to be representative for the Agis population under study. However, as only total expenses are the subject of this study and there appear to be no differences between respondents and non-respondents, it is decided that there is no need to standardize research results in this respect.

Table A3.10: Average expenses in 2001 EUROs per expenses category for respondents and non-respondents. [a]

Expenses category	Respondents	Non-Respondents	T-test	Total
Hospital/rehabilitation [b]	699	748		726
Pharmaceutical drugs [b]	288	252	*	268
Obstetrics [b]	3	4		4
Maternity care [b]	7	6		6
Paramedic care [b]	46	38	*	42
Dental care [c]	12	10	*	11
Medical devices [b]	69	60	*	64
Sick-transport [c]	25	27		26
Total [b]	1,149	1,145		1,147

[a] Effects shown are linearly standardized with respect to age, gender, eligibility, region, pharmaceutical and clinical diagnoses.
[b] The Cochran and Cox/Satterthwaite's approximate two-sided t-test for equality of means between respondents and non-respondents is applied, given that the null hypothesis of equality of variances had to be rejected (Folded form two-sided F-test for equality of variances, $p<0.05$).
[c] The two-sided t-test for equality of means between respondents and non-respondents is applied, given that the null hypothesis of equality of variances could not be rejected at 0.05 confidence level (Folded form two-sided F-test for equality of variances, $0.05<p<0.10$ for dental care and sick-transport).
* Mean expenses of non-respondents differ statistically significantly from that of respondents at 0.05 confidence level.

In Table A3.11 again the focus is on respondents only. Estimation results are presented from a logit model where the probability is modeled for response to the first two mailings versus response to the last two mailings when following the Dillman (1978) four-step mailing procedure. There appears to be no relationship with age, except that for those between 45 and 54 years of age relatively more early respondents than late respondents can be found. The same holds for female respondents. Respondents living in the seventh and tenth regional cluster tend to respond earlier than those living in the second regional cluster. On the other hand, enrollees on social welfare tend to respond later than those employed. Except for those suffering from HIV/Aids, amongst enrollees with pharmaceutical and clinical diagnoses early and late respondents tend to balance out.

In Table A3.12 average health care expenditures of respondents are compared to those of non-respondents for specific categories of health care expenditures (except GP). After indirect standardization (age, gender, eligibility, region, pharmaceutical and clinical diagnoses), it appears that total expenses do not differ statistically significantly between early and late respondents. Health care expenses on pharmaceutical drugs, paramedic care, dental care, and medical devices are

Table A3.11: Odds ratio's and 95% confidence intervals for the administrative variables as explanatory variables for early response versus late response to the survey.

Explanatory variable	Odds ratio	95% confidence interval
15-24	0.95	0.79 - 1.13
25-34	1.00	---
35-44	0.99	0.87 - 1.13
45-54	1.19	1.04 - 1.37
55-64	1.13	0.99 - 1.30
65-74	1.02	0.74 - 1.41
75-84	0.90	0.65 - 1.24
>=85	0.74	0.51 - 1.06
M 15-24	0.75	0.59 - 0.95
M 25-34	0.73	0.62 - 0.86
M 35-44	0.70	0.60 - 0.82
M 45-54	0.72	0.62 - 0.84
M 55-64	0.85	0.74 - 0.98
M 65-74	0.87	0.77 - 0.99
M 75-84	1.03	0.89 - 1.20
M >=85	1.28	0.88 - 1.85
Disabled	0.95	0.86 - 1.04
Employed	1.00	---
Social welfare	0.66	0.58 - 0.76
Unemployed	1.05	0.90 - 1.23
Retired	1.11	0.83 - 1.50
Self-employed	1.13	0.90 - 1.41
APE Region 1	1.02	0.91 - 1.14
APE Region 2	1.00	---
APE Region 3	1.11	1.00 - 1.24
APE Region 4	1.06	0.95 - 1.18
APE Region 5	1.07	0.97 - 1.17
APE Region 6	1.02	0.92 - 1.13
APE Region 7	1.16	1.06 - 1.27
APE Region 8	1.05	0.87 - 1.25
APE Region 9	1.11	0.94 - 1.32
APE Region 10	1.18	1.03 - 1.34
Asthma/COPD	0.97	0.87 - 1.07
Epilepsy	1.20	0.91 - 1.57
Crohn/Colitis Ulcerosa	1.47	0.90 - 2.43
Cardiac disease	0.97	0.87 - 1.08
Rheumatism	1.26	0.84 - 1.89

Explanatory variable	Odds ratio	95% confidence interval
Parkinson	0.99	0.63 - 1.56
Diabetes (Type I)	1.00	0.86 - 1.17
Transplantation	1.57	0.94 - 2.63
Cystic fibrosis	1.19	0.47 - 2.99
Neuromuscular disorder	2.05	0.89 - 4.73
HIV/Aids	6.72	2.38 - 18.96
Renal disease/ESRD	2.17	0.78 - 6.00
DCG1	1.17	0.86 - 1.59
DCG2	0.93	0.71 - 1.22
DCG3	1.32	0.96 - 1.81
DCG4	0.86	0.64 - 1.17
DCG5	0.96	0.67 - 1.38
DCG6	1.01	0.63 - 1.62
DCG7	1.01	0.69 - 1.48
DCG8	0.86	0.57 - 1.31
DCG9	1.32	0.54 - 3.20
DCG10	1.08	0.52 - 2.23
DCG11	0.79	0.37 - 1.71
DCG12	0.72	0.39 - 1.32
DCG13	0.87	0.45 - 1.65
Pseudo-R^2	0.89%	

lower in case of late respondents. As total expenditures are the main focus of this study, no standardization will be conducted in this respect.

The SF-36 scales are only available for respondents. In Table A3.13 the eight SF-36 subscales are included in the logit model of early response versus late response. A mixed pattern is revealed. Amongst the physical health scales, higher PF and GH scores are associated with early response, whereas higher RP scores is associated with later response. Amongst the mental health scales, higher scores on the RE and MH scales are associated with early response, whereas higher scores on the VT scale are associated with later response. In order to get a better grasp of the association with physical an mental health, it is therefore recommended to include the PCS and MCS scale scores instead of the eight SF-36 subscales in the logit model.

In order to get a clearer picture of the influence of physical and mental health on early and late responsiveness, in Table A3.14 results are presented for the logit model with PCS and MCS as explanatory variables. Physical health seems not to influence early or late responsiveness, enrollees in better mental health appear to respond early rather than late.

Table A3.12: Average expenses in 2001 EUROs per expenses category for early and late respondents.

Expenses category	Early respondents	Late respondents	T-test	Total
Hospital/rehabilitation [b]	704	690		699
Pharmaceutical drugs [b]	301	262	*	288
Obstetrics [b]	3	4		3
Maternity care [b]	6	6		7
Paramedic care [b]	48	42	*	46
Dental care [c]	13	11	*	12
Medical devices [b]	71	63	*	69
Sick-transport [c]	26	23		25
Total [b]	1174	1101		1149

[a] The Cochran and Cox/Satterthwaite's approximate two-sided t-test for equality of means between respondents and non-respondents is applied, given that the null hypothesis of equality of variances had to be rejected (Folded form two-sided F-test for equality of variances, $p<0.05$). * Mean expenses of non-respondents differ statistically significantly from that of respondents at 0.05 confidence level.

Table A3.13: Odds ratio's and 95% confidence intervals for the SF-36 scores as explanatory variables for early response versus late response to the survey. [a]

Explanatory variable	Odds ratio	95% confidence interval
PF	1.003	1.002 - 1.005
RP	0.998	0.997 - 0.999
BP	0.999	0.997 - 1.000
GH	1.004	1.001 - 1.006
VT	0.996	0.993 - 0.998
SF	0.999	0.997 - 1.001
RE	1.002	1.001 - 1.003
MH	1.005	1.002 - 1.007
Pseudo-R^2	1.93%	

[a] Effects shown are standardized with respect to age, gender, eligibility, region, pharmaceutical and clinical diagnoses.

Table A3.14: Odds ratio's and 95% confidence intervals for the SF-36 scores as explanatory variables for early response versus late response to the survey. [a]

Explanatory variable	Odds ratio	95% confidence interval
PCS	1.000	0.996 - 1.004
MCS	1.007	1.004 - 1.010
Pseudo-R^2	1.12%	

[a] Effects shown are standardized with respect to age, gender, eligibility, region, pharmaceutical and clinical diagnoses.

To summarize, the analysis of response and nonresponse shows that female enrollees tend to respond relatively more frequent than male enrollees, and relatively most respondents are to be found amongst those between 35 and 84 years of age. There are also some differences in response tendency to be found for subgroups based on eligibility, region of living, and pharmaceutical and clinical diagnoses. However, most importantly, it is concluded that neither 2001 total expenses nor the timing of returning the questionnaire form determine the choice to respond. Given that total expenses is the variable that the focus is on in this research study, because of this result no standardization for selective nonresponse needs to be done. Note that the variables in order to conduct this nonresponse analysis are available for both respondents and non-respondents from the Agis sickness fund administration.

Appendix A3.6: Summary of administrative, survey and external variables

Table A3.15: Items derived from the administrative claims data, the Agis Health Survey 2001, and the APE and Prismant external data sources.

Topics	Number of items		
	Survey	**Claims data**	**Other sources**
Potential Access *- Availability*			
• University and general hospital bed density within 25 kilometers of ZIP-code			1
• Nursery home bed density within 25 kilometers of ZIP-code			1
• General practitioner with own pharmacy		1	
Potential Access *- Organization*			
Potential Access - *Predisposing variables*			
• Sex	1	1	
• Ethnicity	3	1	
• Age	1	1	
• Education	1		
• Marital status	1		
• Household size	1		
• Insurance type		1	
• Region		1	
• Percentage of immigrants per ZIP-code			1
• Percentage of sickness fund members per ZIP-code			1
• Percentage of one-person households per ZIP-code			1
• Percentage of low-income households per ZIP-code			1
• Degree of urbanization per ZIP-code			1
• Less prosperous ZIP-code areas			1

Topics	Number of items		
Potential Access - *Enabling variables*			
• Main wage earner	1		
• Monthly family income after taxes (categorical)	1		
• Supplementary insurance policy		3	
• Distance (in km) to the nearest university or general hospital			1
• Distance (in km) to the nearest home practitioner			1
Potential Access - *Need*			
• Acute diseases and complaints during last 2 months	7		
• Chronic diseases ever	5		
• Chronic diseases in last 12 months	15		
• Chronic diseases identified by pharmaceutical claims data	1	12	
• Chronic diseases identified by diagnostic hospital data	1		13
• Functional disabilities in communication and mobility	8		
• Number of days in bed because of illness or injury during last 6 months	1		
• Anxious or worried for at least 2 weeks consecutively	1		
• Down or depressed for at least 2 weeks consecutively	1		
• Height and weight, to construct a body-mass index	1		
• In need of home care	1		
• Registered on home care waiting list	1		
Realized Access - *Utilization of Health Services*			
• Consultation of a general practitioner during last two months	1		
• Consultation of a specialist during last year	1	1	
• Admission to a hospital or clinic during last year	5	1	
• Number of days spent at hospital or clinic during last year		1	
• Consultation of paramedic therapist during last year	8	1	
• Consultation of ambulatory mental care ever resp. during last year [a]	10	1	
• Consultation of alternative practitioner during last year [a]	7	1	
• Use of home care during last year	2		
• Number of hours of home care during last year	1		
• Use of prescribed drugs during last 2 weeks resp. last year	1	1	
• Use of drugs without a prescription during last 2 weeks	1		
Realized Access - *Consumer Satisfaction*			
• Satisfaction with current amount of home care	1		
Health and Well-Being			
• Physical functioning (PF scale)	10		
• Role-physical (RP scale)	4		
• Bodily pain (BP scale)	2		
• General health perceptions (GH scale)	5		
• Vitality (VT scale)	4		
• Social functioning (SF scale)	2		
• Role-emotional (RE scale)	3		
• Mental health (MH scale)	5		
• Perceived change in health (TRAN)	1		
• Standardized Mortality Ratio per ZIP-code			1
• Standardized Mortality Ratio (age < 75) per ZIP-code			1

[a] These claims data are available with respect to supplementary insurance policies that are offered by Agis.

4 Chapter

HEALTH STATUS MEASUREMENT

A self-administered mail survey mode that includes the SF-36 questionnaire is used to collect information on health status in this study. The choice of a generic instrument is made as a measurement of health status for a general population consisting of people with and people without (non-specific types of) diseases should be obtained. Therefore, health status is measured from the sickness fund members' point of view.

Table 4.1 describes the 35 items that constitute the eight SF-36 scales, as well as a single-item self-evaluated change in health status measure. A full description of the standard SF-36 questionnaire can be found in Ware, Snow and Kosinski (1993, 2000).[77] In Appendix A4.1 completeness and consistency of the SF-36 item responses used in this study are presented.

In section 4.1 the construction of the eight SF-36 physical and mental health scales is established and the assessment of their quality in terms of response, completeness, reliability and validity.

4.1 Construction of eight SF-36 health scales

Ware, Davies-Avery, Brook et al. (1980) mention several reasons in favor of using multi-item measures, amongst others reducing the number of scores necessary to describe the health status dimensions of interest and reducing the number of missing cases by imputing score estimations based on available other items belonging to the same construct. In order to construct the health scales the Likert (1932) rule of summation is applied, i.e. the items are summed up unstandardized and unweighted.[78]

The first necessary condition for applying this summation rule is equality of item means and item variances contributing to the same scale. In Table 4.2 item means and standard deviations are presented for the 19,741 respondents for whom all eight scales could be derived.

---

77. For this study, the official translation into Dutch that followed the stepwise, iterative procedures developed by the IQOLA Project is used (Aaronson et al. 1998). This version distinguishes from an alternative translation carried out by Van der Zee, Sanderman, and Heyink (1996). Although similar, the two translations are not identical. Additionally, the Van der Zee, Sanderman, and Heyink (1996) version follows the scoring algorithms for the RAND-36 rather than for the SF-36. These differences in scoring procedures between the two versions of the questionnaire render the results pertaining to two of the eight scales (BP and GH) non-comparable.

78. When constructing the scales for ten items the score has been reversed, such that for all 36 items a higher item score represents a better health status.

Table 4.1: Abbreviated content for items in each SF-36 scale

Health scale	Item	Q	Item content
Physical Functioning (PF)	PF01	3a	Vigorous activities, such as running, lifting heavy objects, strenuous sports
	PF02	3b	Moderate activities, such as moving a table, vacuuming, bowling
	PF03	3c	Lifting or carrying groceries
	PF04	3d	Climbing several flights of stairs
	PF05	3e	Climbing one flight of stairs
	PF06	3f	Bending, kneeling, or stooping
	PF07	3g	Walking more than a mile
	PF08	3h	Walking several blocks
	PF09	3i	Walking one block
	PF10	3j	Bathing or dressing
Role-Physical (RP)	RP1	4a	Limited in the kind of work or other activities
	RP2	4b	Accomplished less than would like
	RP3	4c	Cut down the amount of time spent on work or other activities
	RP4	4d	Difficulty performing the work or other activities
Bodily Pain (BP)	BP1	7	Intensity of bodily pain
	BP2	8	Extent pain interfered with normal work
General Health (GH)	GH1	1	Is your health: excellent, very good, good, fair, poor
	GH2	11a	I seem to get sick a little easier than other people
	GH3	11b	I am as healthy as anybody I know
	GH4	11c	I expect my health to get worse
	GH5	11d	My health is excellent
Vitality (VT)	VT1	9a	Feel full of pep
	VT2	9e	Have a lot of energy
	VT3	9g	Feel worn out
	VT4	9i	Feel tired
Social Functioning (SF)	SF1	6	Extent health problems interfered with normal social activities
	SF2	10	Frequency health problems interfered with social activities
Role-Emotional (RE)	RE1	5a	Cut down the amount of time spent on work or other activities
	RE2	5b	Accomplished less than one would like
	RE3	5c	Didn’t do work or other activities as carefully as usual

Health scale	Item	Q	Item content
Mental Health (MH)	MH1	9b	Been a very nervous person
	MH2	9c	Felt so down in the dumps nothing could cheer you up
	MH3	9d	Felt calm and peaceful
	MH4	9f	Felt downhearted and blue
	MH5	9h	Been a happy person
Reported Health Transition	TRAN	2	Rating of health now compared to one year ago

Source: Ware and Kosinski (1993, 2000)

Table 4.2 shows that the item means and variances are comparable within the respective scales. Therefore, standardization is not necessary when constructing the scales.[79]

In order to meet the necessary condition of a linear relationship between the item scores and the underlying health construct, one general health (GH) item and one item in the bodily pain (BP) scale have been recalibrated. Recalibration means that the item scores are transformed non-linearly, the transformation being based on NSFHS survey results (Ware, Snow and Kosinski (1993, 2000)). For the remaining 34 items it appears from empirical work that the assumption of a linear relationship between item scores and the underlying health construct is met and therefore recalibration is not necessary.

In order to test the linear relationship between item scores and the constructed health scales the corresponding linear correlations are determined.[80] A comparison is made between the item correlations contributing to the same scale and equality is tested. Items contribute to the scale to the same extent if comparability holds. Items that are not comparable should be deleted from the scale, the remaining items can be included into the scale with equal weights.

Although Table 4.2 shows substantial variation between these correlations, the test outcomes may be judged as positive. This conclusion is justified when all items appear to contribute substantially to the total scale score, i.e. with correlation

79. Although the physical scale items (PF) show an increasing tendency in the average item score, McHorney, Ware, Lu et al. (1994) do not decide against applying the summation method of Likert (1932) for this reason.

80. In order to perform the test of the Likert scaling assumption, the eight scales are not rescaled to the 0-100 spectrum. It would have led to incorrect estimations of the item-scale correlations as the items themselves are not rescaled individually. From a separate analysis it appears that when rescaling nonetheless, the item-scale correlations would have been higher and more distinct from the McHorney et al. (1994, Table 6) correlation estimates.

Table 4.2: Item means, standard deviations and Pearson correlations [a] between SF-36 items and scales [b]

Item	Mean	St.dev.	PF	RP	BP	GH	VT	SF	RE	MH
PF01	1.91	0.78	0.65*	0.55	0.51	0.54	0.43	0.41	0.27	0.25
PF02	2.34	0.73	0.82*	0.64	0.57	0.59	0.50	0.51	0.35	0.31
PF03	2.35	0.73	0.80*	0.62	0.59	0.56	0.48	0.49	0.34	0.31
PF04	2.32	0.76	0.82*	0.57	0.51	0.55	0.46	0.45	0.32	0.28
PF05	2.59	0.64	0.81*	0.55	0.48	0.51	0.42	0.46	0.32	0.26
PF06	2.33	0.74	0.77*	0.56	0.54	0.51	0.42	0.44	0.31	0.27
PF07	2.33	0.80	0.82*	0.58	0.51	0.54	0.45	0.46	0.33	0.27
PF08	2.53	0.71	0.81*	0.53	0.48	0.50	0.42	0.46	0.33	0.26
PF09	2.65	0.62	0.76*	0.49	0.44	0.46	0.38	0.44	0.32	0.25
PF10	2.80	0.47	0.56*	0.40	0.38	0.39	0.34	0.41	0.28	0.24
RP1	1.65	0.48	0.57	0.77*	0.54	0.54	0.51	0.55	0.49	0.37
RP2	1.59	0.49	0.59	0.80*	0.56	0.58	0.55	0.57	0.53	0.41
RP3	1.59	0.49	0.64	0.83*	0.59	0.57	0.52	0.55	0.45	0.36
RP4	1.56	0.50	0.64	0.82*	0.60	0.60	0.56	0.57	0.47	0.39
BP1	4.44	1.38	0.58	0.60	0.87*	0.57	0.53	0.55	0.36	0.40
BP2	4.23	1.41	0.62	0.65	0.87*	0.59	0.56	0.60	0.40	0.42
GH1	3.11	1.01	0.63	0.62	0.60	0.73*	0.62	0.58	0.42	0.47
GH2	3.99	1.18	0.38	0.43	0.39	0.54*	0.47	0.46	0.35	0.41
GH3	3.32	1.25	0.41	0.41	0.38	0.57*	0.44	0.41	0.28	0.34
GH4	3.33	1.20	0.46	0.41	0.39	0.51*	0.40	0.37	0.28	0.30
GH5	3.34	1.36	0.58	0.59	0.56	0.76*	0.62	0.57	0.40	0.48
VT1	4.22	1.38	0.27	0.32	0.29	0.38	0.48*	0.42	0.37	0.52
VT2	3.82	1.40	0.48	0.53	0.48	0.59	0.69*	0.57	0.44	0.58
VT3	4.29	1.26	0.48	0.53	0.52	0.58	0.69*	0.60	0.46	0.57
VT4	3.65	1.25	0.49	0.54	0.53	0.59	0.69*	0.57	0.43	0.53
SF1	4.21	1.08	0.49	0.56	0.53	0.54	0.59	0.71*	0.55	0.58
SF2	3.83	1.18	0.54	0.60	0.56	0.60	0.65	0.71*	0.54	0.62
RE1	1.74	0.44	0.35	0.49	0.35	0.40	0.48	0.54	0.77*	0.55
RE2	1.69	0.46	0.36	0.52	0.37	0.42	0.49	0.54	0.77*	0.55
RE3	1.74	0.44	0.32	0.43	0.33	0.37	0.44	0.49	0.69*	0.49
MH1	4.54	1.17	0.26	0.31	0.31	0.37	0.43	0.45	0.45	0.61*
MH2	5.06	1.12	0.31	0.38	0.37	0.44	0.55	0.58	0.55	0.74*
MH3	4.17	1.32	0.29	0.39	0.38	0.46	0.64	0.55	0.49	0.70*
MH4	4.71	1.14	0.28	0.36	0.36	0.43	0.60	0.56	0.53	0.76*
MH5	4.52	1.28	0.20	0.30	0.29	0.39	0.55	0.48	0.41	0.66*

[a] Item-total correlations, corrected for overlap. Standard error = 0.0071.
[b] N = 19,741 (only respondents for whom all eight scales could be calculated).
[*] Correlation between item and (assumed) corresponding health scale.

coefficients that exceed the 0.40 standard for item-internal consistency (Ware, Snow and Kosinski 1993, 2000). Furthermore, Ware and Gandek (1998) mention that an unequal weighting scheme seldomly leads to an increase in scale applicability, because of scoring complexity.

Table 4.3 shows the distribution of the eight scale scores on the 0-100 spectrum. All scales appear to describe the full scorerange from 0 to 100, but the average scale scores and corresponding standard deviations may differ to a high extent. The reason for this is the difference in number of items between the eight health scales and the resulting difference in number of scoring levels (see the second and third row in Table 4.3).

Table 4.3: Statistical characteristics of the eight SF-36 scales [a]

	PF	RP	BP	GH	VT	SF	RE	MH
Number of items	10	4	2	5	4	2	3	5
Number of response categories	3	2	6/5	5	6	5	2	6
Number of different scale score levels [b]	21	5	10	21	21	9	4	26
Mean	70.82	59.83	66.69	60.45	59.91	75.45	72.25	72.03
Median	80.00	75.00	72.00	62.00	60.00	87.50	100.00	76.00
Minimum	0	0	0	0	0	0	0	0
Maximum	100	100	100	100	100	100	100	100
Range	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100
Standard deviation	28.55	43.62	27.08	23.03	21.21	26.11	39.64	19.52
Coefficient of variation	40.32	72.90	40.61	38.09	35.40	34.60	54.86	27.11
Skewness	-0.83	-0.39	-0.35	-0.37	-0.34	-0.94	-0.97	-0.75
Kurtosis	-0.44	-1.63	-0.82	-0.64	-0.41	0.07	-0.77	0.15
% respondents with scalescore 0	1.66	27.89	1.66	0.42	0.45	1.64	17.98	0.13
% respondents with scalescore 100	19.86	47.96	26.46	2.16	2.30	36.60	62.53	5.54

[a] N = 19,741
[b] In practice there may exist more scale score levels than indicated in this table, because of the imputation of missing responses by the (not rounded-off) arithmetical average of the non-missing responses.

Furthermore, it is apparent that the scale scores are distributed unevenly on all eight scales, in all cases there is an *over*representation of the respondents with a positive health score. The role-emotional (RE) and role-physical (RP) scales show relatively high percentages of respondents with the highest (100) or lowest (0)

score attainable. The social functioning (SF) scale as well shows a substantial percentage of respondents with score 100. It is obvious that the limited number of possible scale levels explains these so-called floor- and ceiling-effects. Although the physical functioning (PF) and bodily pain (BP) scales have more scale levels, there still are ceiling-effects to a certain extent. The general health (GH), vitality (VT) and mental health (MH) scales show no major floor- or ceiling-effects.

4.1.1 Completeness

Table 4.4 shows that the percentage of respondents that have filled out all items appears to be lower than the corresponding survey percentages in the Medical Outcomes Study (MOS) for five out of eight health scales (cf. McHorney, Ware, Lu, et al. 1994).[81] Three health scales appear to perform better with the Agis Health Survey 2001.

Table 4.4: Percentage of respondents with complete items in each SF-36 scale

	PF	RP	BP	GH	VT	SF	RE	MH
Agis SF-36 [a]	83.8	89.1	**96.9**	88.1	**93.8**	93.1	91.8	**94.0**
MOS SF-36 [b]	**88.3**	**93.8**	92.2	**92.5**	93.6	**93.7**	**95.1**	90.3

[a] N = 23,002
[b] N = 3,445

In order to assess the data completeness, Table 4.5 shows the percentage of respondents for whom the health scales could be calculated, after imputation of missing item scores when necessary. With imputation the average of non-missing item responses is substituted for the missing item score, provided that no more than half of the items are missing.

Table 4.5: Percentage of respondents for whom scale scores could be calculated (after imputing missing items when applicable)

	PF	RP	BP	GH	VT	SF	RE	MH
Agis SF-36 [a]	93.6	92.2	98.1	93.9	**97.4**	98.9	93.5	97.0
MOS SF-36 [b]	**96.4**	**96.2**	**99.4**	**97.3**	97.2	**99.6**	**96.0**	**99.1**

[a] N = 23,002
[b] N = 3,445

In order to be conclusive on the completeness of the SF-36 data in the Agis Health Survey, the sample is restricted to those respondents for whom all eight health

81. The Medical Outcomes Study (MOS) was a two-year study of patients with chronic conditions conducted in 1986-1987. The 116-item MOS core survey measures of quality of life include physical, mental, and general health. The SF-36 was developed during the MOS to measure generic health concepts relevant across age, disease, and treatment groups.

scales are available (again, after imputing missing items when applicable). It appears that for 85% of the respondents on the Agis Health Survey all eight health scales can be derived (19,741 respondents). The test results described above make way to apply the Likert (1932) scoring algorithm. The eight SF-36 scales constructed this way should now be tested for reliability and validity. Validity and reliability are commonly applied criteria to determine measurement quality of a survey. A survey is called valid if the survey items constitute a measurement of the construct intended to be measured, in our case the constructs "physical health" and "mental health". A survey is called reliable when the measurement results can be reproduced in repetitive instances.

Scale validity is determined in order to answer the question whether the observed scale scores may indeed be interpreted in terms of the underlying health construct one wants to measure. Van den Brink and Mellenbergh (1998, p. 59) show that a reliable scale score does not necessarily have to be valid, i.e. reliability is a necessary but not sufficient condition for validity. In other words, measurement errors of chance can be minimal, but that does not imply that the scale adequately measures the health construct under study. In the next two sections the eight SF-36 health scales are tested for reliability and validity.[82]

4.1.2 Reliability

In order to apply the SF-36 scores as health scales, they have to be reliable and valid. As health scales are based on empirical evaluations all kinds of errors can cause the empirical score to deviate from the true, latent score. Reliability is a measure of the extent to which empirical measures capture these latent scores and is commonly expressed in terms of internal consistency by Cronbach's alpha.

Reliability is a measure for the precision with which the observed health scores can describe population differences in the latent health construct. In order to determine reliability the latent health scores are unknown, however. A way around this problem is to start from the Nunnally (1967) proposition that measurement errors caused by respondents not reading the questions correctly and by lessened attention amongst others, are the most important sources of differences between observed and latent testscores. Under this assumption, in classical test theory the so-called reliability coefficient is then derived as follows.

Suppose the following linear relationship holds between the latent, true score T and the observed, empirical test score X:

$$X = T + \varepsilon$$

82. The results presented in these sections are not corrected for population differences.

where ε describes the measurement error and thus a certain amount of error in the answers to the test. The test is called more reliable to the extent that *X* and *T* show a stronger association. The coefficient of reliability ρ_{XT}^2 is the reliability measure and is the squared statistical correlation ρ_{XT} between *X* and *T*. The next equality can be derived (e.g. Van den Brink and Mellenbergh 1998):

$$\rho_{XT}^2 = \sigma_T^2 / \sigma_X^2 = \sigma_T^2 / (\sigma_T^2 + \sigma_\varepsilon^2)$$

where σ_T^2 is the latent, theoretical variance and σ_X^2 the variance of the observed scale scores.

Notice that σ_T^2 cannot be observed, but σ_X^2 can be calculated easily given the available test scores. Different methods exist to overcome the problem of unobserved σ_T^2 when determining the coefficient of reliability ρ_{XT}^2. A distinction can be made between methods based on repeated observations and those based on only one measurement observation. The Cronbach (1951) method is most applied in surveys in order to estimate the reliability coefficient, also with the SF-36 test construction (McHorney, Ware, Lu, et al. (1994)). According to this method a so-called coefficient alpha α can be estimated, which defines a lower bound on the reliability coefficient ρ_{XT}^2:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_X^2}\right)$$

where *k* is the number of scale items and σ_i^2 the sample variance in item *i*. It should be noted that with methods based on one measurement observation, reliability is defined in terms of internal consistency.[83]

From Table 4.6 all scale scores appear to stand the tests, as for these group comparisons the reliability coefficient appears to be greater than the threshold of 0.70 of the Nunnally (1978) rule.[84]

In Table 4.6 scale homogeneity is also presented, i.e. the average inter-item correlation per health scale. A higher coefficient of reliability can merely result from a vast number of items *k* that is needed to construct a scale, although these items do not have much in common (in terms of inter-item covariances σ_{ij}). This heterogeneity then results in a low homogeneity score, because this is independent

83. This can be seen by noticing that $\sigma_X^2 = \sum_{i=1}^{k} \sigma_i^2 + (\sum \sum_{i \neq j} \sigma_{ij})$, such that $\alpha = \frac{k}{k-1}(\sum \sum_{i \neq j} \sigma_{ij} / \sigma_X^2)$.

84. For individual level comparisons this minimum equals 0.90 (Nunnally 1978). The Helmsteader (1964) rule of 0.50 is an alternative to the minimum 0.70 that applies to group comparisons.

Table 4.6: SF-36 scale homogeneity and internal-consistency reliability [a]

Scale	k [b]	Homogeneity [c]	Reliability [d]
PF	10	0.62	0.94
RP	4	0.73	0.91
BP	2	0.87	0.93
GH	5	0.49	0.82
VT	4	0.53	0.81
SF	2	0.71	0.83
RE	3	0.68	0.86
MH	5	0.57	0.87

[a] N = 19,741.
[b] Number of items and number of item-internal consistency test per SF-36 scale.
[c] Average inter-item correlation.
[d] Internal-consistency reliability (Cronbach's alpha).

from the number of scale items *k*, in contrast to Cronbach's alpha α. From Table 4.6 the average inter-item correlations appear acceptable. Note that length of the SF-36 scales in general is seen as limited and thus its potential diffusion.

4.1.3 Validity

Validity has to do with the question whether it is reasonable to interpret test results in terms of the underlying concept that is supposed to be measured. In the literature there appear several definitions of the concept validity. According to Van den Brink and Mellenbergh (1998) they all can be categorized in two main categories: construct validity and criterion oriented validity. With construct validity the central idea is to find support for adequate coverage of the underlying construct, criterion oriented validity is about determining the extent to which predictions of an external criterion can be derived.

Aaronson, Muller, Cohen et al. (1998) tested the validity of the Dutch standard form SF-36 scales by one-way analysis of variance with respect to the variables self-reported age, sex and chronic conditions, both from an Amsterdam sample (1994, *N* = 4,172) and a national sample (1996, *N* = 1,742). One-way analysis of variance is also used to test the variables self-reported occurrence and number of migraine attacks in a sample of clinically identified migraine sufferers (1993, *N* = 423, standard SF-36 form). Furthermore, it is concluded that the Dutch language version of the SF-36 is a valid instrument for use in both general population surveys and in studies of chronic disease populations in the Netherlands. Note that these validity tests are all based on self-reported indicators.[85]

85. The validity of the Dutch *acute form* is tested in Aaronson, Muller, Cohen et al. (1998) by a one-way analysis of variance of SF-36 scales subdivided by applying objective (clinical) measures of

The main contribution of the validity tests with respect to the standard form SF-36 scales derived in this section is that (1) sample size of the database used here is substantially bigger than those used by Aaronson, Muller, Cohen et al. (1998), and (2) both self-reported and more objective, administrative variables derived from the claims administration are used in order to check the validity of the SF-36 scales.

Construct validity

With construct validity it is determined whether the scale measure describes the observable variable in a representative way, and whether the composition of the measure is adequate. In general it is about the question whether one truly measures one supposes to measure.

As an example, construct validity plays a crucial role with tests on the progress of education. An examination for a certain subject must relate to that subject and the questions should be posed such that the examination mark represents the level of knowledge adequately. Furthermore, McHorney, Ware and Raczek (1993) showed a very large difference in average scores at the MOS mental health scale between patients with a relatively minor and uncomplicated medical condition and patients with a psychiatric illness. As psychiatric patients have poor mental health by definition of their disease, this significant average lower score is a demonstration of construct validity for the Mental Health scale.

Construct validity thus relates to the coverage of the content domain and the relationship with other constructs and tests. With respect to the relationship with other constructs and tests a distinction can be made between internal and discriminant validity. In the validating process, both are required. It has to be shown that a test does indeed correlate with other measures of the same construct (convergent validity) and does not correlate with other non-related constructs (discriminant, also called divergent validity). A high extent of both convergent and discriminant validity give support to the construct validity of a test.

Content validity

Content validity is determined by the extent to which the survey represents the construct universe. Content validity requires the existence of a defining standard against which one can compare the content of a measure. Standards can be based on well-accepted theoretical definitions, on published standards, or on interviews with those who are experiencing the types of health problems under study. In

---

disease, i.e. clinical disease stage and Karnofsky Performance Status in a sample of cancer patients who started either a new course of chemotherapy or radiotherapy (1992-1994, N = 485).

constructing the SF-36, Ware (1987) describes some standards for evaluating the content validity of general health measures intended to be comprehensive. Ware, Snow and Kosinski (1993, 2000) give a description of the lowest and highest scale scores for each SF-36 scale. For ease of reference, this is shown in Table 4.7 below.

From Table 4.7 it appears that the most and least positive answers are associated with good and bad health respectively. Content-based descriptions of intermediate scores are also useful to understand the meaning of differences in scale scores between the extremes by zooming in on its underlying item scores. With respect to the physical functioning scale (PF), Figure 4.1 shows the relationship between the scale score and the percentage of respondents that is able to "walk one block or more" (physical functioning item PF07). This figure reveals an increasing number of respondents able to walk one block as scale scores are higher, conform the theoretical expectations. This is an example of a content-based interpretation of the intermediate scale scores. For all other nine physical activities an analogous conclusion can be drawn (these figures are not shown here).

In Table 4.8 for ten general health (GH) score classes the percentage of respondents is shown that evaluate their current health (general health item GH1)

Table 4.7: Content-based descriptions of lowest and highest scale scores

	Meaning of scores	
Concepts	**Lowest possible score**	**Highest possible score**
Physical functioning	Limited a lot in performing all physical activities including bathing or dressing due to health	Performs all types of physical activities including the most vigorous without limitations due to health
Role-physical	Problems with work or other daily activities as a result of physical health	No problems with work or other daily activities as a result of physical health
Bodily pain	Very severe and extremely limiting pain	No limitations due to pain
General health	Evaluates personal health as poor and believes it is likely to get worse	Evaluates personal health as excellent
Vitality	Feels tired and worn out all of the time	Feels full of pep and energy all of the time
Social functioning	Extreme and frequent interference with normal social activities due to physical or emotional problems	Performs normal social activities without interference due to physical or emotional problems
Role-emotional	Problems with work or other daily activities as a result of emotional problems	No problems with work or other daily activities as a result of emotional problems
Mental health	Feelings of nervousness and depression all of the time	Feels peaceful, happy, and calm all of the time

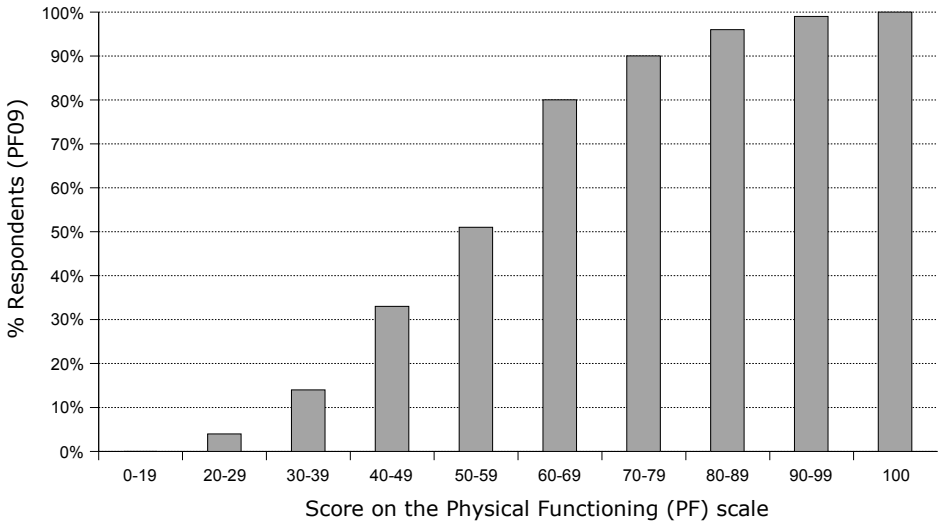


Figure 4.1: The percentage of respondents that can walk one block or more (physical functioning item PF09), given the score on the physical functioning scale

as "excellent" (response=1), "good" (response=3) respectively "fair" or "poor" (response =4 or response =5). As an example, 31.0% of the respondents with general health (GH) score between 91 and 100 evaluate their health as "excellent", 27.6% as "good" and 0.0% as "fair" of "poor". At the bottom of the general health (GH) scale all respondents evaluate their health as "fair" or "poor". In between

Table 4.8: Percentage of respondents that evaluate their health as "excellent", "good", and "fair" or "poor".

Levels	General health (GH) scale		Percentage of respondents [a]		
	Average score	Number of respondents	"Excellent"	"Good"	"Fair" or "Poor"
91-100	95.2	1849	31.0%	27.6%	0.0%
81-90	84.8	2549	14.2%	52.4%	0.5%
71-80	74.5	3354	5.2%	76.6%	1.7%
61-70	64.9	3101	1.9%	85.1%	6.6%
51-60	55.4	2304	1.0%	70.7%	25.8%
41-50	46.3	1830	0.2%	30.8%	68.6%
31-40	37.1	1895	0.0%	9.6%	90.3%
21-30	27.8	1394	0.0%	2.2%	97.8%
11-20	17.9	848	0.0%	0.2%	99.8%
0-10	6.8	450	0.0%	0.0%	100.0%
Total	60.4	19574	6.1%	48.3%	33.2%

[a] Responses to the question: In general, would you say your health is? (general health item GH1)

Table 4.9: Percentage of respondents that reporting feeling tired (vitality scale item VT4) and having a lot of energy (vitality scale item VT2) for ten levels of the vitality scale (VT).

Levels	Vitality scale (VT)		Percentage of respondents	
	Average score	Number of respondents	Tired [a]	Lot of energy [a]
91-100	97.5	902	0.0%	100.0%
81- 90	87.0	1890	0.0%	96.1%
71- 80	77.3	3425	0.3%	79.9%
61- 70	67.6	3346	1.4%	42.6%
51- 60	57.6	3006	5.4%	16.9%
41- 50	47.7	2772	15.7%	6.3%
31- 40	37.8	2081	38.6%	2.8%
21- 30	27.8	1217	78.6%	1.0%
11-20	18.0	586	95.6%	0.0%
0- 10	6.0	363	99.7%	0.0%
Total	59.9	19588	17.0%	39.0%

[a] All or most of the time, in the past 4 weeks

these extremes, a theoretically sound pattern in the respondents' evaluations appears.

With respect to the vitality (VT) scale, for each ten-point class Table 4.9 shows the percentage of respondents that have been feeling tired (vitality scale item VT4) all or most of the time, or have a lot of energy (vitality scale item VT2) all or most of the time. Here as well there appears a response pattern consistent with theoretical expectations: the percentage of respondents with a lot of energy (vitality scale item VT2) is highest at the highest vitality (VT) scale levels and the percentage feeling tired (vitality scale item VT4) is lowest. At the lowest scale levels the pattern is reversed.

Convergent and discriminant validity

Campbell and Fiske (1959) test content validity in terms of convergent and discriminant validity based on the so-called multitrait-multimethodmatrix. A "trait" represents the construct one tries to measure, the "method" represents the way the test data are collected. In this study, the multitrait-multimethod matrix is as shown in Table 4.2, i.e. the correlationmatrix between the SF-36 scales ("trait") and the items ("method"). Based on this matrix the test results for convergent validity and discriminant validity are determined, which are shown in Table 4.10.

In case of item-convergent validity the item is sufficiently linearly correlated with the underlying concept one supposes to measure. This test consists of calculating the correlation between an item and its corresponding scale exclusive of that item. In this way these correlations are corrected for the contribution of the item to its corresponding scale. This means 10 tests for the physical functioning

Table 4.10: Item scaling tests for SF-36 scales [a]

Scale	k [b]	Convergent validity [c]	Convergent validity test [d]	Discriminant validity [e]	Discriminant validity test [f]
PF	10	0.56-0.82	10/10	0.24-0.64	70/70
RP	4	0.77-0.83	4/4	0.36-0.64	28/28
BP	2	0.87	2/2	0.36-0.65	14/14
GH	5	0.51-0.76	5/5	0.28-0.63	35/35
VT	4	0.48-0.69	4/4	0.27-0.60	27/28
SF	2	0.71	2/2	0.49-0.65	14/14
RE	3	0.69-0.77	3/3	0.32-0.55	21/21
MH	5	0.61-0.76	5/5	0.20-0.64	35/35

[a] N = 19,741.

[b] Number of items and convergent validity tests per SF-36 scale.

[c] Correlations between items and corresponding SF-36 scale corrected for overlap.

[d] Scaling success convergent validity = Number of correlations significantly larger than 0.40/ total number of correlations (corrected for overlap).

[e] Correlations between items and other SF-36 scales than the corresponding scale.

[f] Scaling success discriminant validity = Number of occasions in which the item correlation with the corresponding scale is significantly larger (>= 2 standard deviations) than the correlation with the other scale/total number of correlations.

scale (PF) should be evaluated, for example. Item convergent validity holds if the item-scale correlation is at least 0.40 (after correcting for item-scale overlap). From Table 4.10 it appears that the item-convergent validity test is passed for each of the eight scales.

The test for item-discriminant validity consists of determining whether the correlation for an item with the scale it is part of, is significantly larger than with the other scales. In this respect, the number of tests to be performed equals the number of items in each scale multiplied by the total number of scales constructed minus one. For example, for the physical scale 10 x (8-1) = 70 tests should be performed. Only in case of the vitality-scale one of the tests performed for item-discriminant validity shows that an item correlates significantly larger with other scales than with the vitality-scale itself (0.52 resp. 0.48). In McHorney, Ware, Lu, et al. (1994, Table 7) all tests show positive results, in Aaronson, Muller, Cohen et al. (1998, Table 2) there was also only one negative test result for the vitality (VT) scale in the general population sample. The conclusion must be that in principle the corresponding item should not be included in the vitality scale, as it does not contribute enough to the interpretation of the scale. Nonetheless, following Aaronson, Muller, Cohen et al. (1998), the item is maintained as an underlying part of the vitality scale, in order to concur with international scale constructions.

Criterion-related validity

Criterion-related validity is also called predictive validity. When the criterion-related validity of a measure should be determined, this means that the predictive value for external behavior or external events is estimated. As an example, the Dutch CITO-examination at the end of primary school is supposed to have predictive power for future educational performance. In this case it is less important that the CITO-examination describes the current level of knowledge (construct validity) than functioning as a predictor of future educational performance.

Thus these tests concern the relationship between the measure and external criteria independent from the ones in question. These external criteria can be sampled concurrently or in the future. The CITO-examination is an example of an external criterion of the latter type. A clinical criterion like sickness or disease severity is an example of an external criterion of the former type.[86] In this study, the criterion-related validity is determined by linking the SF-36 scale scores to information on sickness, disease and utilization measures.

In Ware, Snow and Kosinski (1993, 2000) criterion-related validity is tested for six out of eight SF-36 health scales plus the health change item. These criteria are chosen because (1) they are of clinical and/or social importance, (2) they constitute plausible outcomes of variation in functioning and well-being measured by the scales, and (3) they are independent on the scale under review. Here analogous results are presented with respect to the Agis Health Survey 2001.

In general, it is expected to be more difficult to do paid work in case of physical limitations. Table 4.11 presents the relationship between the **physical functioning (PF) scale** and the percentage of respondents that is disabled to work in a paid job. The figures in the table apply to the labor market population only, i.e. those respondents that are (self-)employed, unemployed and/or disabled. Between the highest and the lowest scale levels, a nearly perfect ordering of the percentages of people that cannot work appears. Respondents with lower scale scores are more often disabled than others. Thus, these percentages signify a social interpretation to the observed differences in scale scores. To a somewhat smaller extent, the same pattern holds for age. Age may be a cause for and/or a confounder of disability in the relationship with the physical functioning (PF) scale.

---

86. Indicators of physical and mental health that are the result of (medical) expert opinions can be applied as concurrent external criteria. Given a certain disease, McHorney, Ware and Raczek (1993) rank-order the patients according to severity and determine the correlation with these rankings. Although this kind of indicators is frequently understood as being objective, it is frequently observed that medical experts differ a lot in their opinions about the severity of a specific disease.

Table 4.11: Percentage of respondents that cannot work because of health problems and their mean age, for ten levels of the physical functioning (PF) scale. [a]

Levels	Physical functioning (PF) scale		Percentage cannot work	Average age
	Mean score	Number of respondents		
91-100	98.2	5263	6.7%	38.4
81-90	87.8	1909	17.7%	43.7
71-80	77.6	1140	30.7%	45.6
61-70	67.4	895	40.2%	47.0
51-60	57.4	659	45.6%	46.9
41-50	47.5	572	51.9%	48.1
31-40	37.3	437	57.9%	49.9
21-30	27.4	340	68.2%	48.8
11-20	17.3	260	66.6%	50.0
0-10	5.6	235	68.7%	50.6
Total	79.4	11710	17.9%	42.8

[a] Only respondents who are (self)employed, unemployed and/or disabled in 2001, i.e. exclusive of those on social welfare and pensioners.

The criterion-related validity of the **role-physical (RP) scale** is determined by cross-examining the average score with the general health (GH) scale. It may be expected that respondents evaluate their general health differently for each of the five role-physical scale levels. From Table 4.12 it is apparent that the average general health (GH) score is indeed significantly different for the respondents in the five role-physical (RP) classes distinguished ($F = 3397.51$, $p < 0.0001$).

Furthermore, the last column of Table 4.12 is shown in order to get an answer to the question whether the role-physical (RP) scale can be interpreted as an interval scale. Therefore the average general health (GH) scores in the last column but one have been transformed to a 0-100 scale. These transformed general health (GH) scores appear to be comparable with the role-physical (RP) scores presented in the first column.

In order to test the criterion-related validity of the **bodily pain (BP) scale** the same criterion is applied as was the case with the physical functioning scale. Table 4.13 presents the percentage of respondents who have a paid job. Again, the figures in the table apply to the labor market population only, i.e. those respondents that are (self-)employed, unemployed and/or disabled. The ordering of the percentages of people is less perfect than in Table 4.11, but still the discrepancies between the highest and the lowest scale levels are distinct. Respondents with lower scale scores are more often disabled than others. Thus, these percentages signify a social interpretation to the observed differences in scale scores. The age pattern seems to be somewhat less distinct, as compared to the results presented in Table 4.11.

Table 4.12: Mean general health (GH) scores for respondents at five levels of the role-physical (RP) scale a)

Scores (RP)	Number of respondents	Percentage of respondents	General Health (GH) evaluation	
			Mean score	Transformation to 0-100 scale
100	9337	48.4%	74.5	100.0
75	1530	7.9%	62.8	66.0
50	1444	7.5%	56.5	47.2
25	1682	8.7%	48.8	24.6
0	5309	27.5%	40.4	0.0
Total	19302	100.0%	60.6	59.2

a) From an ANOVA test it appears that average general health (GH) scores differ significantly between the five role-physical (RP) levels ($F = 3397.51$, $p < 0.0001$).

Table 4.13: Percentage of respondents not being able to work for ten levels of the bodily pain (BP) scale. [a]

Levels	Bodily pain (BP) scale		Percentage cannot work	Average age
	Mean score	Number of respondents		
91-100	100.0	3477	11.4%	40.8
81-90	84.0	1424	9.7%	39.5
71-80	73.7	1626	15.5%	41.4
61-70	62.0	1759	22.6%	43.1
51-60	51.3	870	33.0%	44.6
41-50	41.1	1227	50.1%	48.0
31-40	31.4	498	52.9%	47.2
21-30	22.1	502	57.8%	46.8
11-20	12.6	121	63.0%	46.6
0-10	2.4	206	53.1%	46.8
Total	70.0	11710	18.4%	42.8

[a] Only respondents who are (self)employed, unemployed and/or disabled in 2001, i.e. exclusive of those on social welfare and pensioners.

Tests for the criterion-related validity of the **general health (GH) scale** can be performed by cross-linking these score levels with the percentage of respondents that were hospitalized the past 12 months, the annual rate of visits to the general practitioner and the number of prescriptions per GP visit. For each of these variables it may be expected that higher values of these external criteria correspond to higher general health scores. Table 4.14 shows that the expected associations do indeed exist. It should be noted that in other tables the general health (GH) scale acted as an indication of the criterion-related validity of the other scales. In those cases, it is assumed that the general health (GH) scores can be interpreted as a direct measure of someone's personal health situation. Table 4.14 confirms this hypothesis.

Table 4.14: Health care utilization rates for respondents differing in general health evaluations (general health item GH1).

General health item 1 (GH1)	General health (GH) scale score		Hospitalized in the past 12 months (in general or academic institution)	Contacted a general practitioner in the past 12 months	Used prescribed medicines in the past 14 days
	Average score	Trans-formation to 0-100 scale			
Excellent	90	100	3.9%	55.6%	21.4%
Very good	83	90	5.2%	69.3%	31.6%
Good	68	69	7.8%	79.4%	59.3%
Fair	38	26	16.9%	92.0%	87.5%
Poor	20	0	26.7%	92.9%	90.9%
Total	61	59	10.7%	80.9%	63.1%

Table 4.15: Mean mental (MH) health scores of respondents for four levels of the role-emotional (RE) scale.

Score (RE)	Number of respondents	Percentage of respondents	Mental health (MH) score	
			Average score	Transformation to 0-100 scale
100	12,344	62.8%	80.3	100.0
66.7	1,871	9.5%	69.4	62.7
33.3	1,906	9.7%	60.0	30.5
0	3,549	18.0%	51.1	0.0
Total	19,670	100.0%	72.1	71.9

Note: An ANOVA test shows that the mean mental health (MH) scores differ significantly across the four role-emotional (RE) levels ($F = 3680.98$, $p < 0.0001$).

The criterion-related validity of the **role-emotional (RE) scale** has been determined by crossing it with the mental (MH) health scale scores. From Table 4.15 it appears that these scores differ significantly across the four role-emotional levels ($F = 3680.98$, $p < 0.0001$). Furthermore, after transformation of these scores to a 0-100 scale it shows that the role-emotional (RE) scale can be interpreted as an interval scale.

The **mental health (MH) scale** is tested by crossing these scores against mental health care utilization numbers. Table 4.16 shows that 55.6% of respondents with zero mental health (MH) score (i.e. minimal mental health) contacted an ambulant mental care institution (RIAGG), 22.2% contacted social workers, a psychiatrist or psychotherapist, and 11.1% frequented the psychiatric outpatient clinic. For those scoring 100 on the mental health (MH) scale (i.e. maximum mental health) these percentages are smaller than one. In general, respondents appear to contact social workers and institutions more frequently as their mental health score is lower.

Table 4.16: Percentage of respondents that contacted social workers or institutions for six selected mental health (MH) scores

Social workers or institutions	N	Score = 0	Score = 20	Score = 40	Score = 60	Score = 80	Score = 100
Psychiatric (dpt. of) hospital	16	0.0%	5.5%	0.0%	0.6%	0.4%	0.4%
Ambulant mental care (RIAGG)	218	55.6%	34.5%	20.9%	9.1%	5.0%	0.9%
Crisis center	27	0.0%	10.9%	2.8%	1.2%	0.4%	0.0%
General social work	180	22.2%	20.0%	16.9%	6.5%	4.7%	1.5%
Psychologist	123	0.0%	7.3%	6.3%	6.7%	3.6%	0.7%
Psychiatrist	69	22.2%	10.9%	4.7%	3.2%	1.6%	0.4%
Psychotherapist (excl. psychiatrist)	62	22.2%	3.6%	5.1%	2.9%	1.5%	0.4%
Psychiatric outpatient clinic	55	11.1%	14.5%	3.5%	1.9%	1.5%	0.3%
Alcohol/drugs consultation office	15	0.0%	3.6%	0.0%	1.2%	0.4%	0.0%
Sexuologist	17	0.0%	0.0%	0.8%	0.7%	0.6%	0.1%
Another psychosocial caretaker	43	0.0%	0.0%	4.7%	2.7%	0.6%	0.3%

4.2 Conclusions

In this chapter, the extent to which the eight constructed SF-36 scales can readily be applied in our study as measures of physical and mental health differences is examined. As the assumptions underlying the summation method of Likert (1932) were tested positively, this scoring algorithm is applied to construct the eight SF-36 scales (physical functioning, role-physical, bodily pain, general health, vitality, social functioning, role-emotional and mental health).

The completeness, reliability and validity of the eight constructed SF-36 scales for the survey sample at hand are tested following the guidelines set by Ware, Snow and Kosinski (1993, 2000) and Ware, Kosinski and Keller (1994). The results of this chapter are summarized in Table 4.17, where the results of the Agis Health Survey 2001 are compared to those of the MOS and IQOLA surveys.[87] In terms of sample size, the Agis Health Survey 2001 scores best.

Compared to the studies in the IQOLA project, response rate and response completeness in the Agis Health Survey 2001 appears to be at the lower end. The completeness score of 85% is comparable to the figures with respect to Norway,

87. The "International Quality of Life Assessment" (IQOLA) project aims at translating, validating and norming the SF-36 health survey for use in multinational "clinical trials" and other international studies. In 1998 experiences from eleven countries have been evaluated (see Ware, J.E., "Editorial", Journal of Clinical Epidemiology, Volume 51, issue 11, pp. 891-892). At the time of writing, the SF-36 has been translated in more than 40 languages.

Table 4.17: Information about survey methods, data quality and respondents: the Agis survey, the MOS survey and the IQOLA general population surveys

	Survey methods				Data quality			Respondents		
Country	Sampling Frame	Mode of administration	Year of administration	Response rate (%)	Data completeness [a] (%)	Scaling success rate [b] (%)	Range of reliability [c]	Sample size	% Male	Mean age (SD)
Denmark	Sample drawn from Central Person Registry data	SF-36 self-administered after a structured personal interview and returned by mail	1994	68	88	99.3	.76-.92	4084	48	43.9 (17.8)
France	National sample drawn from Sofres METASCOPE database	Mail survey	1995	81	93	99.3	.79-.91	3656	48	44.6 (18.1)
Germany	National sample drawn from Infratest database	SF-36 self-administered after a structured personal interview and handed to interviewer	1994	61	95	98.2	.74-.94	2914	48	45.2 (18.4)
Italy	National sample drawn from electoral lists	Self-administered and returned to interviewer (50%) or interviewer-administered (50%)	1995	NA [d]	99	100.0	.77-.93	2031	49	47.7 (17.1)
Japan	National sample drawn from government registration lists	Self-administered and collected by trained data collector	1995	75	99	96.8	.68-.87	3395	49	43.4 (17.1)
Netherlands (National)	National sample drawn from national telephone registry	Mail survey	1996	63	96	98.9	.77-.92	1771	56	47.6 (18.0)

	Survey methods				Data quality			Respondents		
Country	**Sampling Frame**	**Mode of administration**	**Year of administration**	**Response rate (%)**	**Data completeness [a] (%)**	**Scaling success rate [b] (%)**	**Range of reliability [c]**	**Sample size**	**% Male**	**Mean age (SD)**
Netherlands (Amsterdam)	Sample drawn from Amsterdam municipal population registry	SF-36 self-administered after a structured personal interview and handed to interviewer	1994	50	95	99.6	.75-.93	4059	46	43.1 (18.1)
Netherlands (Agis Health Insurance)	**Sample drawn from Agis health plan membership administration**	**Mail survey**	**2001**	**46**	**85**	**99.6**	**.81-.94**	**23163**	**42**	**55.2 (18.4)**
Norway	Mail survey	Mail survey	1996	67	86	99.6	.79-.90	2323	49	44.9 (16.5)
Spain	National sample from Institutio Nacional de Estadistica	Interviewer administration	1996	80	98	99.6	.77-.96	9151	48	45.2 (18.6)
Sweden	Seven community samples from various regions of Sweden	Mail survey	1991-1992	68	89	100.0	.79-.91	8930	48	42.6 (16.6)
United Kingdom (National)	National sample drawn from Office of Population Census and Surveys Omnibus survey	Interviewer administration	1992	78	99	100.0	.81-.93	2056	48	41.3 (18.6)
United Kingdom (Sheffield)	Sample drawn from two general practice lists in Sheffield	Mail survey	1991	83	93	98.6	.76-.93	1582	45	41.3 (15.4)
United States of America (National - NSFHS)	National sample drawn from 1989 and 1990 General Social Surveys	Mail (68%) or telephone (32%) survey	1990	77	97	98.2	.68-.93	2474	48	43.6 (17.4)

	Survey methods				Data quality			Respondents		
Country	Sampling Frame	Mode of administration	Year of admini-stration	Response rate (%)	Data complete-ness [a] (%)	Scaling success rate [b] (%)	Range of reli-ability [c]	Sample size	% Male	Mean age (SD)
United States of America (Boston, Chicago, Los Angeles - MOS)	Sample drawn from HMO, multispecialty groups and solo practices in three cities	SF-36 self-administered after initial survey and a structured personal interview and returned by mail	1986-1987	71	96	100.0	.78-.93	3445	38	- (-)

[a] Source: Agis Health Survey 2001, McHorney, Ware, Lu, et al. (1994) and Gandek and Ware (1998).

[b] Percent of respondents who completed at least 50% of items for all scales.

[c] Percent of tests in which items correlated significantly higher with hypothesized scales than with competing scales (out of 280 tests).

[d] Cronbach's alpha.

[e] NA = Not Applicable: Due to the method of sampling, the response rate is not an appropriate indicator of participation.

Denmark and Sweden. As a potential explanation, it is observed that completeness is also relatively low for those surveys in the IQOLA project that are conducted by mail as compared to other administration modes. Another explanation might be that the average age of the respondents to the Agis Health Survey is above 50, whereas it is below 50 for all surveys included in the IQOLA project.

Reliability and validity of the eight SF-36 scales has been tested positively in this chapter. Table 4.17 shows that scaling success rate and range of reliability for the Agis Health Survey 2001 are at the higher end as compared to the IQOLA studies.

Finally, it should be noted that in many applications the physical component scale (PCS) and mental component scale (MCS) are used as two summary measures instead of the eight separate health scales. Although there is a loss of information, Ware, Kosinski, Bayliss et al. (1995) mention that there are theoretically important advantages such as an increase in the number of scoring levels, elimination of floor and ceiling effects, smaller intervals of confidence around the scale scores, and the simplicity of using two instead of eight scales. Nonetheless, given our study design, the eight scales are preferred to the two summary scales. In this way, the advantages mentioned are kept and the drawback of loosing part of the available scale information is avoided.[88]

88. For the interested reader, in Appendix A4.3 the PCS and MCS factor score coefficients are presented as well as the correlations between the SF-36 scales and rotated principal components, both for the Agis and the general U.S. population (NSFHS).

Appendix A4.1: The SF-36 item responses

In this section first a description is given of the response quality with respect to the 36 items of the SF-36 questionnaire. Next completeness and response consistency of these items is determined.

Response

The more answers are missing, the less confidence one has in the scales derived from these items. In Table 4.2 for each item the percentage of missing observations is calculated for 23,002 respondents, as well as the mean, standard deviation and frequency distribution of the item results. These are the original item values and not transformed in any way. [89]

The percentage of missing responses as reported in Table A4.1 is higher on average than the corresponding percentage in McHorney, Ware, Lu et al. (1994).

Completeness

From Table A4.1 it appears that all applicable item response categories have been filled out. Had this not been the case, it could have pointed to interpretation problems with respect to those response categories. Furthermore, the item relative frequency distribution appears to be comparable to the results shown in Table 4.3 in McHorney, Ware, Lu, et al. (1994).

Response consistency

Data quality can also be expressed in terms of response consistency, i.e. to what extent do item responses contradict each other.

As an example, there appears to be a response inconsistency if a respondent is able to walk a mile or more without limitations (physical functioning item PF08) but one block is too much (physical functioning item PF09). With respect to the SF-36 questionnaire, Ware, Snow and Kosinski (1993, 2000) apply fifteen internal consistency checks based on pairs of SF-36 items which constitute the so-called Response Consistency Index (RCI).[90] In Table A4.2 the RCI for the Agis respondents is compared with the RCI for the general US population (NSFHS) and for the MOS patients with one or more selected chronic diseases. The data quality for the Agis sample seems comparable to the NSFHS and MOS results.

---

89. There have been 23,163 out of 50,022 insurees that have returned a non-empty form, 23,002 of them have filled out at least one out of 36 items. In this chapter the focus is on these 23,002 respondents who filled out one or more items.

90. See Appendix A4.2.

Table A4.1: Item means, standard deviations and relative frequency distribution [a]

Item				Item relative frequency distribution							
Name	**Label**	**Mean**	**SD**	**Missing**	**1**	**2**	**3**	**4**	**5**	**6**	**Total** [c]
PF01	Vigourous	1.89	0.78	7%	34%	35%	24%	- [b]	-	-	100%
PF02	Moderate	2.32	0.75	7%	16%	32%	45%	-	-	-	100%
PF03	Lifting	2.32	0.74	7%	16%	32%	46%	-	-	-	100%
PF04	ClimbingSeveral	2.30	0.77	8%	18%	29%	45%	-	-	-	100%
PF05	ClimbingOne	2.57	0.66	9%	9%	22%	61%	-	-	-	100%
PF06	Bending	2.30	0.75	7%	17%	32%	45%	-	-	-	100%
PF07	WalkingMile	2.31	0.82	8%	21%	22%	49%	-	-	-	100%
PF08	WalkingSeveral	2.51	0.73	9%	13%	19%	60%	-	-	-	100%
PF09	WalkingOne	2.64	0.64	9%	8%	17%	66%	-	-	-	100%
PF10	Bathing	2.79	0.49	7%	3%	13%	77%	-	-	-	100%
RP1	Limited	1.64	0.48	8%	33%	59%	-	-	-	-	100%
RP2	Accomplished	1.58	0.49	8%	39%	53%	-	-	-	-	100%
RP3	Cut down	1.59	0.49	9%	38%	53%	-	-	-	-	100%
RP4	Difficulty	1.55	0.50	7%	42%	51%	-	-	-	-	100%
BP1	Pain	2.75	1.41	2%	27%	17%	21%	23%	8%	2%	100%
BP2	Limited	2.06	1.14	3%	40%	27%	17%	9%	4%	-	100%
GH1	CurrentHealth	3.17	0.91	5%	6%	11%	45%	28%	5%	-	100%
GH2	Easier	3.97	1.19	6%	4%	8%	21%	15%	46%	-	100%
GH3	Equal	2.69	1.25	7%	19%	26%	24%	14%	10%	-	100%
GH4	Worse	3.29	1.20	6%	8%	13%	41%	10%	23%	-	100%
GH5	Excellent	2.67	1.37	6%	20%	34%	10%	16%	14%	-	100%

Item				Item relative frequency distribution							
Name	**Label**	**Mean**	**SD**	**Missing**	**1**	**2**	**3**	**4**	**5**	**6**	**Total** [c]
VT1	Full of pep	2.79	1.39	4%	17%	34%	12%	21%	7%	4%	100%
VT2	Energy	3.20	1.42	4%	11%	26%	16%	25%	14%	5%	100%
VT3	Worn out	4.27	1.28	4%	3%	6%	14%	32%	22%	19%	100%
VT4	Tired	3.63	1.27	3%	8%	9%	20%	39%	14%	6%	100%
SF1	Extent	1.81	1.10	3%	52%	24%	11%	6%	4%	-	100%
SF2	Frequency	3.79	1.19	5%	4%	10%	25%	19%	37%	-	100%
RE1	Cut down	1.73	0.44	6%	25%	69%	-	-	-	-	100%
RE2	Accomplished	1.68	0.47	6%	30%	64%	-	-	-	-	100%
RE3	Carefully	1.73	0.44	6%	25%	69%	-	-	-	-	100%
MH1	Nervous	4.51	1.18	3%	3%	4%	8%	32%	29%	22%	100%
MH2	Down	5.04	1.14	4%	1%	3%	5%	19%	24%	45%	100%
MH3	Peaceful	2.84	1.33	3%	13%	37%	14%	21%	8%	3%	100%
MH4	Downhearted	4.70	1.15	4%	1%	3%	7%	29%	26%	29%	100%
MH5	Happy	3.09	0.81	3%	23%	37%	12%	17%	5%	2%	100%

[a] N = 23,002

[b] Non-applicable response categories

[c] Rounding off cell percentages may have led to these not summing up to 100% in de last column.

Table A4.2: Relative distribution of the SF-36 Response Consistency Index

Number of response inconsistencies	United States (NSFHS) [b,d]	The Netherlands (Agis) [a,d]	United States (MOS) [c]
0	90.3	92.3	94.5
1	6.1	4.3	3.4
2	1.3	1.4	1.1
3	0.8	0.8	0.3
4	0.6	0.9	0.5
5	0.2	0.1	0.1
6	0.4	0.2	0.1
7	0.1	0.0	0.0
8	0.1	0.1	0.0
9-15	0.0	0.0	0.0
Total	100.0	100.0	100.0

[a] N = 23,002

[b] N = 2,474, National Survey of Functional Health Status (NSFHS), 1990 (Ware, Snow and Kosinski (1993, 2000))

[c] N = 3,434, Medical Outcomes Study (MOS), 1986-1987 (Ware, Snow and Kosinski (1993, 2000))

[d] By rounding off the row percentages they may not sum up to 100.0 percent.

Appendix A4.2: Response Consistency Index

The purpose of this functionality is to describe the scoring of the Response Consistency Index (RCI) for the SF-36 Version 1.0. (Source: http://www.qualitymetric.com/misc/Publications.aspx).

Significance

This functionality is not required for scoring the SF-36 scales and summary measures. The SF-36 Response Consistency Index (RCI) is scored to provide users with an index for evaluating the quality of responses to individual survey forms. The RCI consists of fifteen internal consistency checks based on pairs of SF-36 items. For example, a report of being able to "walk one mile" but not "one block" is considered an inconsistency in scoring the RCI.

Algorithm #11: SF-36 Response Consistency Index (RCI)

Algorithm #11 is used to score the RCI. The RCI consists of 15 internal consistency checks based on the following pairs of SF-36 items:

1) item 3i (walking one block) and item 3a (vigorous activities) - (pf09 / pf01)
2) item 3j (bathing or dressing) and item 3a (vigorous activities) - (pf10 / pf01)
3) item 3i (walking one block) and item 3b (moderate activities) - (pf09 / pf02)
4) item 3j (bathing or dressing) and item 3b (moderate activities) - (pf10 / pf02)

5) item 3i (walking one block) and item 3d (climbing several flights of stairs) - (pf09/ pf04)
6) item 3j (bathing or dressing) and item 3d (climbing several flights of stairs) - (pf10 / pf04)
7) item 3i (walking one block) and item 3g (walking more than one mile) - (pf09/ pf07)
8) item 3j (bathing or dressing) and item 3g (walking more than one mile) - (pf10/ pf07)
9) item 9d (felt calm and peaceful) and item 9b (been a very nervous person) - (mh3 / mh1)
10) item 9h (been a happy person) and item 9f (felt downhearted and blue) - (mh5 / mh4)
11) item 9e (have a lot of energy) and item 9g (feel worn out) - (vt2 / vt3)
12) item 9a (feel full of pep) and item 9i (feel tired) - (vt1 / vt4)
13) item 1 (in general, would you say your health is) and item 11d (my health is excellent) - (gh1 / gh5)
14) item 7 (how much bodily pain) and item 8 (how much did pain interfere) - (bp1 / bp2)
15) item 6 (extent health interferes with social act.) and item 10 (time health interferes w/social act.) - (sf1 / sf2)

Scoring the RCI consists of assigning a value of 1 to each inconsistent response and a value of 0 to each consistent response to the 15 pairs of items. A total RCI score is computed by summing across the values (1/0) assigned to the 15 pairs of items. Scores on the RCI range from 0 to 15. A total score of 15 indicates that the individual elicited 15 inconsistent responses (poor data quality). A total score of 0 indicates that the individual elicited 0 inconsistent responses (excellent data quality).

Preconditions

To calculate the RCI, values for item responses must be in the appropriate range. Out-of-range values are assumed to be missing, typically denoted as a negative one (-1). The RCI is scored using "precoded" item response values. The RCI will be inappropriately scored if "recoded" item response values are used.

It is not necessary that a respondent has complete data on all 15 internal consistency checks to compute the RCI. If a respondent has incomplete data on any of the items that make up the 15 internal consistency checks, simply add up the values assigned to the internal consistency checks that the respondent has

complete data. A missing score on the RCI is assigned when a respondent has incomplete data on all items that make up the 15 internal consistency checks.

Post-conditions

The RCI score must be an integer value that ranges from 0 to 15. The RCI is not presented as part of the SF-36 health profile. At the individual survey level, a score from 0 to 15 is reported for the RCI. For example, an RCI score of 1indicates that the respondent elicited one inconsistent response among the 15 pairs of items. At the population level, the percentage of the population at each level of the RCI is reported. For example, 90% of the general U.S. population has a score of 0 (no inconsistencies) on the RCI.

Inputs

The input data for the RCI are integers. The integer values are the "precoded" item values for the 20 SF-36 items that make up the 15 internal consistency checks.

Processing

Scoring the RCI included the following steps:

1) RCI1: If the response value for item 3i is 1 (limited a lot) and the response value for item 3a is 2 (limited a little) or 3 (not limited) then a value of 1 is assigned to the item pair; otherwise a 0 is assigned.
2) RCI2: If the response value for item 3i is 1 (limited a lot) and the response value for item 3b is 2 (limited a little) or 3 (not limited) then a value of 1 is assigned to the item pair; otherwise a 0 is assigned.
3) RCI3: If the response value for item 3i is 1 (limited a lot) and the response value for item 3d is 2 (limited a little) or 3 (not limited) then a value of 1 is assigned to the item pair; otherwise a 0 is assigned.
4) RCI4: If the response value for item 3i is 1 (limited a lot) and the response value for item 3g is 2 (limited a little) or 3 (not limited) then a value of 1 is assigned to the item pair; otherwise a 0 is assigned.
5) RCI5: If the response value for item 3j is 1 (limited a lot) and the response value for item 3a is 2 (limited a little) or 3 (not limited) then a value of 1 is assigned to the item pair; otherwise a 0 is assigned.
6) RCI6: If the response value for item 3j is 1 (limited a lot) and the response value for item 3b is 2 (limited a little) or 3 (not limited) then a value of 1 is assigned to the item pair; otherwise a 0 is assigned.
7) RCI7: If the response value for item 3j is 1 (limited a lot) and the response value for item 3d is 2 (limited a little) or 3 (not limited) then a value of 1 is assigned to the item pair; otherwise a 0 is assigned.

8) RCI8: If the response value for item 3j is 1 (limited a lot) and the response value for item 3g is 2 (limited a little) or 3 (not limited) then a value of 1 is assigned to the item pair; otherwise a 0 is assigned.
9) RCI9: If the response value for item 9h is 1 (all the time) and the response value for item 9f is 1 (all the time) then a value of 1 is assigned to the item pair; If the response value for item 9h is 6 (none of the time) and the response value for item 9f is 6 (none of the time) then a value of 1 is assigned to the item pair; otherwise a 0 is assigned;
10) RCI10: If the response value for item 9d is 1 (all the time) and the response value for item 9b is 1 (all the time) then a value of 1 is assigned to the item pair; If the response value for item 9d is 6 (none of the time) and the response value for item 9b is 6 (none of the time) then a value of 1 is assigned to the item pair; otherwise a 0 is assigned;
11) RCI11: If the response value for item 9a is 1 (all the time) and the response value for item 9i is 1 (all the time) then a value of 1 is assigned to the item pair; If the response value for item 9a is 6 (none of the time) and the response value for item 9i is 6 (none of the time) then a value of 1 is assigned to the item pair; otherwise a 0 is assigned;
12) RCI12: If the response value for item 9e is 1 (all the time) and the response value for item 9g is 1 (all the time) then a value of 1 is assigned to the item pair; If the response value for item 9e is 6 (none of the time) and the response value for item 9g is 6 (none of the time) then a value of 1 is assigned to the item pair; otherwise a 0 is assigned;
13) RCI13: If the response value for item 1 is 1 (excellent) and the response value for item 11d is 5 (definitely false) then a value of 1 is assigned to the item pair; If the response value for item 1 is 5 (poor) and the response value for item 11d is 1 (definitely true) then a value of 1 is assigned to the item pair; otherwise a 0 is assigned;
14) RCI14: If the response value for item 6 is 1 (not at all) and the response value for item 10 is 1 (all the time) then a value of 1 is assigned to the item pair; If the response value for item 6 is 5 (extremely) and the response value for item 10 is 5 (none of the time) then a value of 1 is assigned to the item pair; otherwise a 0 is assigned;
15) RCI15: If the response value for item 7 is 1 (none) and the response value for item 8 is 5 (extremely) then a value of 1 is assigned to the item pair; If the response value for item 7 is 6 (very severe) and the response value for item 8 is 1 (not at all) then a value of 1 is assigned to the item pair; otherwise a 0 is assigned.

A total Response Consistency Index score is computed by summing the values (1/0) assigned to the 15 pairs of items:

RCI = (RCI1 + RCI2 + RCI3 + RCI4 + RCI5 + RCI6 + RCI7 + RCI8 + RCI9 + RCI10 + RCI11 + RCI12 + RCI13 + RCI14 + RCI15)

Outputs

The output consists of an integer that ranges from 0 to 15. A score of 0 indicates no inconsistent response sets to the 15 internal consistency checks (excellent data quality). A score of 15 indicates 15 inconsistent responses to the 15 internal consistency checks (poor data quality). A missing value is assigned if the respondent has missing data on all 15 internal consistency checks. Table A4.3 presents the frequency distribution for the Response Consistency Index in the general U.S. Population.

Table A4.3: Frequency Distribution for the SF-36 Response Consistency Index in the General U.S. Population

Number of Inconsistent Responses	Frequency	Percent	Cumulative Percent
0	2234	90.3	90.3
1	152	6.1	96.4
2	32	1.3	97.7
3	19	0.8	98.5
4	16	0.6	99.2
5	6	0.2	99.4
6	11	0.3	99.7
7	2	0.1	99.8
8	1	0.1	99.9
9	1	0.1	100.0
10-15	0	0.0	100.0

Appendix A4.3: The SF-36 Physical and Mental Component Scales

Table A4.4: Agis and general U.S. population factor score coefficients used to derive PCS and MC scores

	Agis (N=19,741)		U.S. Population (N=2,474)	
	PCS	**MCS**	**PCS**	**MCS**
PF	0.44622	-0.26540	0.42402	-0.22999
RP	0.30722	-0.09193	0.35119	-0.12329
BP	0.35220	-0.15782	0.31754	-0.09731
GH	0.23948	-0.01826	0.24954	-0.01571
VT	0.00060	0.25079	0.02877	0.23534
SF	0.02284	0.22837	-0.00753	0.26876
RE	-0.18696	0.41798	-0.19206	0.43407
MH	-0.28056	0.52900	-0.22069	0.48581

Source (U.S. Population): Ware, Kosinski and Keller (1994)

Table A4.5: Correlations between SF-36 scales and rotated principal components in the Agis study and the general U.S. population

	Agis (N=19,741)			U.S. Population (N=2,474)		
	PCS	**MCS**	**h^2/r_{tt}** [a]	**PCS**	**MCS**	**h^2/r_{tt}**
PF	0.87	0.16	0.84	0.85	0.12	0.78
RP	0.78	0.36	0.82	0.81	0.27	0.82
BP	0.79	0.27	0.76	0.76	0.28	0.72
GH	0.72	0.43	0.85	0.69	0.37	0.78
VT	0.50	0.69	0.90	0.47	0.64	0.75
SF	0.53	0.67	0.89	0.42	0.67	0.92
RE	0.25	0.78	0.77	0.17	0.78	0.78
MH	0.17	0.90	0.96	0.17	0.87	0.92
Reliable variance [b]			84.7%			81.5%

Source (U.S. Population): Ware, Kosinski and Keller (1994)

[a] h^2/r_{tt} = Variance in each SF-36 scale explained by the two principal components (h^2) divided by the reliability of each SF-36 scale (r_{tt}).

[b] Percent of the total reliable variance in SF-36 scales explained by the two principal components.

Chapter 5

A DERIVATION OF NORMATIVE COSTS

In Chapters 2, 3 and 4 a motivation is given for the choice of variables from administrative and survey data sources to include as S-type adjusters in the normative equation (2.3). In Section 5.1, the S-type adjusters are described statistically and cross-tabulated in order to check whether expected relationships between their 2001 values and costs observed in 2002 hold for the data used in this study. After this validity check, the normative costs are derived following equation (2.4) in Section 5.2. These normative costs form the basis for the analysis in subsequent chapters.

5.1 A STATISTICAL DESCRIPTION OF THE S-TYPE ADJUSTERS

In this study it is assumed that Dutch government desires premium subsidies that are appropriately adjusted for the S-type risk factors age, gender and health status. In practice, the premium subsidies are adjusted for the REF adjusters which are included as independent variables in the REF equation (2.1). Based on the normative equation (2.3), the premium subsidies are adjusted for the S-type adjusters chosen in this study. Figure 5.1 illustrates the differences between the selection of the REF adjusters and the S-type adjusters. Although age and gender

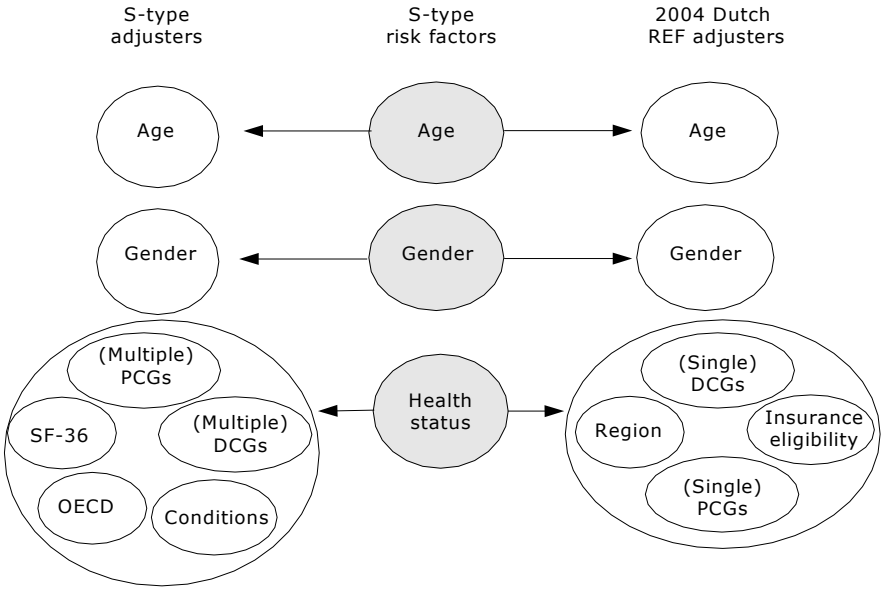


Figure 5.1: The S-type risk factors age, gender and health status and their measures: the S-type adjusters in the normative equation (2.3) and the REF adjusters in the 2004 Dutch REF equation (2.1).

Table 5.1: Administrative and survey measures of the risk factors assumed to be chosen by the Dutch government to be included in the normative regression (2.3)

Dimensions	Measures	Metric	Reference category
Age and gender	Age and gender	Discrete interaction variable, 2 x 8 classes.	Males between 15-24 years
Claims-derived chronic conditions	12 PCGs and 13 DCGs [a,b]	Discrete yes/no variables	PCGs: No PCGs DCGs: No DCGs
Self-reported health status	PF, RP, BP, GH, VT, SF, RE, and MH scales	Continuous variables, 0-100 scale (0=bad health, 100=good health)	None
Functional health status	Number of OECD limitations	Discrete variable, four classes: 0, 1, 2, and 3+ conditions	Zero OECD limitations
Self-reported chronic conditions	Number of chronic conditions	Discrete variable, four classes 0, 1, 2, and 3+ conditions	Zero self-reported conditions diseases

[a] PCG01=Asthma/COPD, PCG02=Epilepsy, PCG03=Crohn/Colitis Ulcerosa, PCG04=Cardiac disease, PCG05=Rheumatism, PCG06=Parkinson, PCG07=Diabetes (Type I), PCG08=Transplantation, PCG09=Cystic fibrosis, PCG10=Neuromuscular disorder, PCG11=HIV/Aids, PCG12=Renal disease/ESRD.

[b] See Appendix A for a description of the DCG classification used in this study.

are included as independent variables both in the REF equation and the normative equation, the implementation of the S-type risk factor health status differs in this study.

In Table 5.1 the set of variables from administrative and survey data sources to be included in the normative regression (2.3) is shown. The S-type adjusters age and gender, PCGs and DCGs, and the self-reported SF-36 health status scales have a continuous metric, functional health status and self-reported chronic conditions have a discrete metric.[91]

It should be noted that in the context of the normative equation an insured can be assigned to multiple pharmacy-based and multiple diagnostic cost groups (i.e. PCGs and DCGs). The PCGs and DCGs are originally constructed in the context of the REF equation such that an insured can be assigned to the single most costly PCG and/or the single most costly DCG. The procedure to construct these cost groups starts with determining the extent of the average difference between observed costs and expected costs for enrollees assigned to such cost groups. Expected costs are costs that might be expected given age and gender and follow from a linear regression of 2002 costs on age and gender interaction terms for those

91. Missing values in case of the survey question on the OECD limitations are recoded to the sample average within the age and sex group to which an enrollee belongs. Missing values in case of the question on self-reported chronic conditions are set to zero by default.

enrollees who do not belong to any cost group. The rank-ordering of cost groups is achieved by applying an iterative procedure where in each iteration (1) the cost group with the largest difference between actual and expected costs is identified, (2) all insured people who belong to this cost group are removed from the data set before going to the next iteration, and (3) the average difference between observed costs and expected costs is recalculated for the remaining enrollees. The order of removing the cost groups in each iteration is the ranking according to decreasing difference between actual and expected costs. This procedure is repeated until the rank of every cost group is determined.[92] The rank-ordering of PCGs and DCGs in the REF equation is applied to mitigate the incentives for strategic upcoding behavior by providers in case of existing comorbidities. This concern is not relevant in the context of the normative equation, because the cross-subsidies will not be based on the S-type adjusters in practice.

Table 5.2 gives a statistical description of these S-type adjusters as well as observed costs, classified by gender and age. Theoretically, if the chosen variables are valid measures of the S-type risk factors, a negative relationship of age with both the Physical Component Scale (PCS) and Mental Component Scale (MCS) is expected, and a positive relationship of age with both the number of OECD limitations and the number of chronic diseases.[93] For reasons of brevity, the PCS and MCS summary scales are presented instead of the eight underlying scales PF, RP, BP, GH, VT, SF, RE, and MH.

From Table 5.2 it appears that morbidity and observed costs increase with age, as expected. Specifically, PCS scores decrease by age, while the opposite holds for the number of OECD limitations and self-reported chronic diseases, both for men and women. MCS scores show a rather mixed pattern, however. It also appears that health care costs in 2002 increase with age, as expected, but drop when people are 85 or older. The explanation may be institutionalization of those enrollees, their health care needs being partly financed out of the Exceptional Medical Expenses Act (AWBZ). To sum up, it may be concluded that the relationships

92. See also Lamers and Van Vliet (2003) for a description of this iterative procedure in the context of the construction of the pharmacy-based cost groups (PCGs).

93. Note that this may only be concluded if the assumption of measurement equivalence holds, i.e. if the health status measurement scales are invariant across subgroups. In that case group differences in the health status measures are proportional to the mean differences in the latent construct (i.e. health status). McHorney, Ware, and Raczek (1993), Ware, Kosinski and Keller (1994) and Ware et al. (1995) prove measurement validity by showing that the SF-36 scales and the PCS and MCS summary scales have a comparable interpretation across subgroups known to differ in severity of physical and/or mental clinical condition.

Table 5.2: PCS, MCS, number of OECD limitations, number of chronic diseases 2001 and observed costs 2002, classified by 2001 gender and age.

Gender / Age	PCS	MCS	Nr. OECD limitations	Nr. chronic diseases	Observed costs 2002
	Mean	Mean	Mean	Mean	Mean
M 15-24	57.25 *	53.11 *	0.06 *	0.02 *	818 *
M 25-34	57.44 *	51.25 *	0.08 *	0.02 *	691 *
M 35-44	55.02 *	50.73	0.22 *	0.03 *	761 *
M 45-54	53.68	50.74	0.33 *	0.10	1901
M 55-64	51.48 *	51.98 *	0.49 *	0.24 *	2378 *
M 65-74	50.39 *	53.13 *	0.49 *	0.34 *	3785 *
M 75-84	47.92 *	51.89 *	0.85 *	0.36 *	5045 *
M >=85	45.96 *	52.61	1.35 *	0.38 *	3742 *
F 15-24	56.21 *	49.20 *	0.10 *	0.01 *	812 *
F 25-34	55.98 *	49.54 *	0.14 *	0.01 *	1256 *
F 35-44	55.01 *	49.84 *	0.21 *	0.03 *	1010 *
F 45-54	53.01	49.75 *	0.46 *	0.08 *	1450 *
F 55-64	51.46 *	51.09	0.52 *	0.17 *	1867
F 65-74	49.11 *	51.18	0.79 *	0.25 *	2814 *
F 75-84	44.60 *	50.53	1.45 *	0.29 *	4262 *
F >=85	41.44 *	49.72	2.60 *	0.31 *	3629 *
Total	53.28	50.69	0.40	0.11	1753

* Statistically significantly different from overall mean (two-sided t-test, p <= 0.05).

between the chosen S-type adjusters, observed costs and age and gender are as expected.

Age and gender are also included as REF adjusters in the 2004 Dutch REF equation. The age classification to be applied in this study, however, differs from the classification applied there. The reasons for this are that (1) survey respondents in this study are at least 16 years of age and (2) with a 10 year age classification the sample distribution of costs appears to resemble the population distribution more closely than with the conventional 5 year age classification. This is due to the limited sample size of the data used in this study.

The closer resemblance between the sample and population distribution of costs in case of a 10 year classification can be observed by comparing Figure 5.2 (5 year classes) with Figure 5.3 (10 year classes). In Figures 5.2 and 5.3 the population distribution of 2002 costs (dotted lines) across age categories is graphed against the sample distribution (straight lines), both for men and women. Several discontinuities exist in Figure 5.2 for men up to the age category 50-54, and for

Men

Women

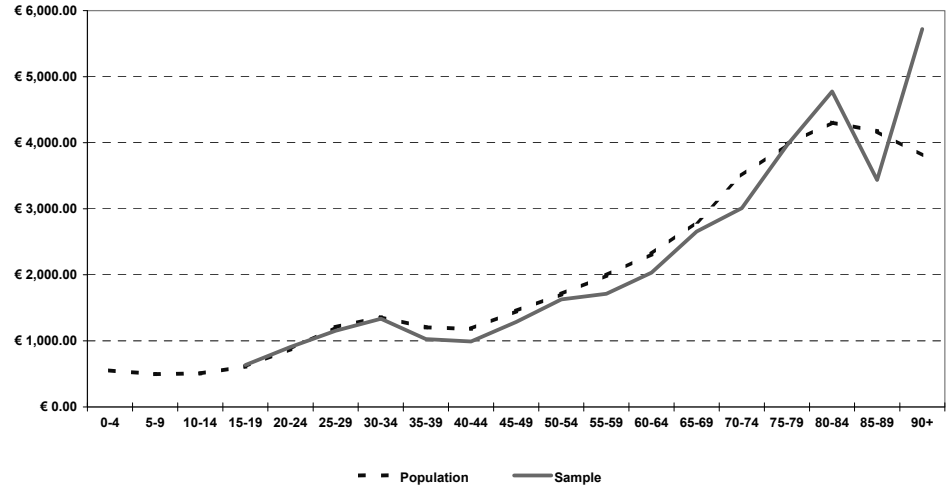
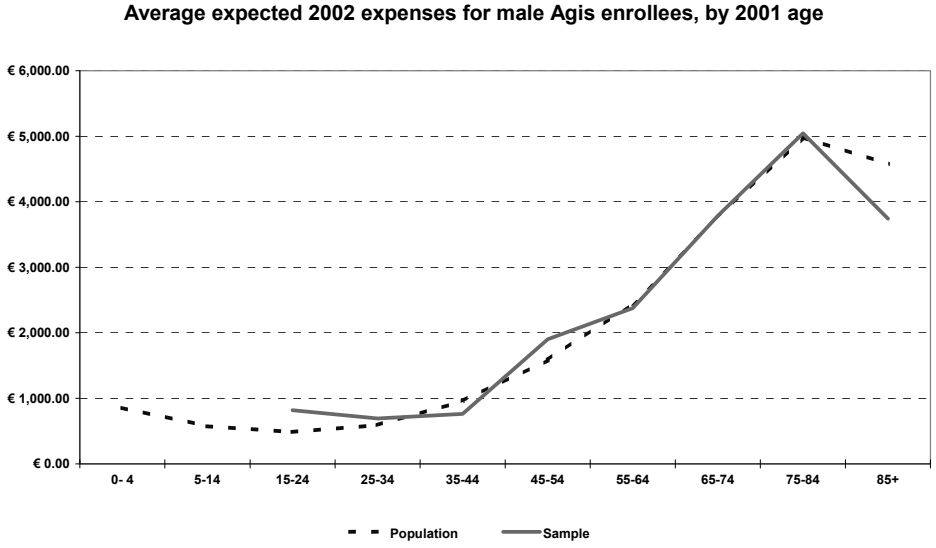


Figure 5.2: Age and gender regression coefficients 2001 obtained from a regression (without other adjusters) of both Agis population and Agis sample 2002 costs, given a five years age classification as in the 2004 Dutch risk-adjustment scheme.

women between 85-89. In Figure 5.3, given that 10 year classes are applied, these discontinuities have disappeared such that the sample pattern more closely follows the population pattern in observed costs 2002.

Men



Women

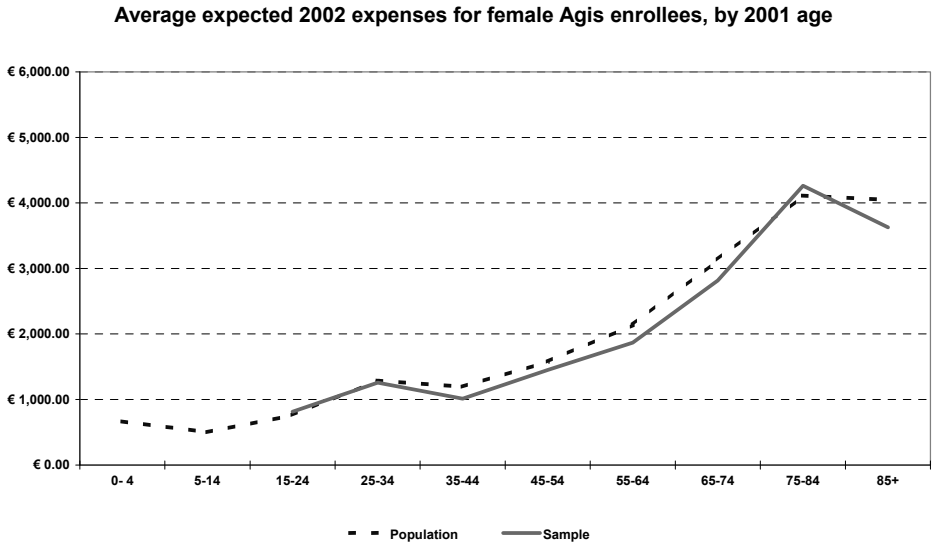


Figure 5.3: Age and gender regression coefficients 2001 obtained from a regression (without other adjusters) of both Agis population and Agis sample 2002 costs, given an alternative ten years age classification.

The research sample is meant to be representative for the population of Agis members and not necessarily for the total Dutch sickness fund population. For comparison purposes, Table 5.3 presents the 2001 Agis and 2004 Dutch national PCG and DCG prevalences given that an insured can belong to a single PCG and

Table 5.3: Number and prevalence of enrollees by 2001 (single) PCG and (single) DCG classes, that are rank-ordered according to the Dutch 2004 classifications.

	Agis [a]			NL [b]
	Unweighted number of enrollees	**Weighted number of enrollees**	**Weighted prevalence per 1000 enrollees**	**Prevalence per 1000 enrollees**
No PCG	14941	16976	911.87	930.30
PCG01	1110	633	34.01	26.66
PCG02	169	90	4.81	4.01
PCG03	52	32	1.71	1.21
PCG04	1377	525	28.17	23.98
PCG05	95	57	3.05	1.42
PCG06	62	23	1.21	0.97
PCG07	640	229	12.32	9.62
PCG08	69	16	0.86	0.79
PCG09	19	7	0.35	0.29
PCG10	28	11	0.61	0.40
PCG11	35	12	0.62	0.30
PCG12	20	8	0.41	0.04
No DCG	17662	18101	972.29	978.12
DCG01	148	78	4.18	3.23
DCG02	140	87	4.66	3.22
DCG03	117	74	3.99	3.48
DCG04	121	78	4.21	3.04
DCG05	93	53	2.86	2.17
DCG06	55	21	1.12	0.99
DCG07	87	45	2.42	2.30
DCG08	68	28	1.50	1.06
DCG09	18	9	0.49	0.40
DCG10	26	10	0.54	0.45
DCG11	20	10	0.55	0.56
DCG12	27	13	0.72	0.54
DCG13	35	9	0.48	0.46

Note: PCG01=Asthma/COPD, PCG02=Epilepsy, PCG03=Crohn/Colitis Ulcerosa, PCG04=Cardiac disease, PCG05=Rheumatism, PCG06=Parkinson, PCG07=Diabetes (Type I), PCG08=Transplantation, PCG09=Cystic fibrosis, PCG10=Neuromuscular disorder, PCG11=HIV/Aids, PCG12=Renal disease/ESRD. See Appendix A for a description of the DCG classification used in this study. An insured can belong to a single PCG and/or a single DCG only.

[a] PCGs and DCGs are based on 2001 data, corrected for the 2002 number of months enrolled.

[b] The national population prevalences are based on Van Vliet, R.C.J.A., and F.J. Prinsze (2004), Table 2 in Section VII (PCGs, 2000 data) and on Van Vliet, Goudriaan and Thio (2003), Tables 6 and 7 in Section I (DCGs, 2001 data).

Table 5.4: PCS, MCS, number of OECD limitations, number of chronic diseases in 2001 and observed costs 2002, classified by 2001 (single) PCG and linearly standardized by 2001 gender and age.[a]

	PCS		MCS		Nr. OECD limitations		Nr. chronic diseases		Observed costs 2002	
PCG	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
No PCG	53.88 *	0.06	50.88	0.07	0.34 *	0.01	0.08 *	0.00	1458 *	45
PCG01	47.71 *	0.32	48.99 *	0.38	0.76 *	0.04	0.16	0.01	3776 *	233
PCG02	49.23 *	0.84	47.57 *	1.00	0.79 *	0.10	0.20 *	0.04	4152 *	615
PCG03	49.26 *	1.40	49.73	1.67	0.16	0.16	0.13	0.06	3213	1031
PCG04	47.02 *	0.36	48.59 *	0.43	1.42 *	0.04	0.70 *	0.01	4750 *	266
PCG05	39.92 *	1.05	51.56	1.25	0.88 *	0.12	0.08	0.05	5885 *	773
PCG06	45.16 *	1.67	46.90	1.99	1.30 *	0.19	0.34	0.07	4985 *	1226
PCG07	47.27 *	0.52	48.52 *	0.63	1.19 *	0.06	1.12 *	0.02	5522 *	386
PCG08	43.47 *	1.98	49.07	2.36	0.67	0.23	0.46 *	0.09	11648 *	1455
PCG09	49.65	3.09	48.24	3.69	0.69	0.36	0.33	0.13	7256 *	2272
PCG10	36.00 *	2.35	47.86	2.81	2.65 *	0.27	0.17	0.10	11631 *	1730
PCG11	50.26	2.33	37.40 *	2.79	0.26	0.27	0.06	0.10	13934 *	1717
PCG12	44.25 *	2.87	51.13	3.43	0.79	0.33	0.62 *	0.12	39498 *	2112
Total	53.28	0.06	50.69	0.07	0.40	0.01	0.11	0.00	1753	44

Note: PCG01=Asthma/COPD, PCG02=Epilepsy, PCG03=Crohn/Colitis Ulcerosa, PCG04=Cardiac disease, PCG05=Rheumatism, PCG06=Parkinson, PCG07=Diabetes (Type I), PCG08=Transplantation, PCG09=Cystic fibrosis, PCG10=Neuromuscular disorder, PCG11=HIV/Aids, PCG12=Renal disease/ESRD. See Appendix A for a description of the DCG classification used in this study. An insured can belong to a single PCG and/or a single DCG only.

* Statistically significantly different from overall mean (two-sided t-test, p <= 0.05).

[a] Enrollees can have only one PCG indication.

a single DCG only.[94] The prevalences with respect to the "No PCG" and "No DCG" subgroups of enrollees show that in our study 2.43% more enrollees belong to a PCG and/or DCG than do nation wide. Relative patterns in the prevalences of the subgroups defined by the PCGs and DCGs appear to be comparable between the Agis population and the national population, however.

PCGs and DCGs are included as S-type adjusters in the normative equation because the association between the SF-36 scores and observed costs might be influenced by treatment effects. As already indicated in Section 2.3, the SF-36

94. Agis population prevalences have not been reported here as these can be seen as strategic business information. Furthermore, in case of the DCGs, Agis population prevalences couldn't even be determined, as they were available for the limited research sample in this study only. Differences between tabulated national prevalences and the Agis sample prevalences may therefore originate from differences between the Agis sample and population prevalences, and/or from differences between the Agis population and national population prevalences.

scores for people under treatment for a chronic disease may be similar to those for people without any disease, although the level of health care expenditures differs between these groups. Furthermore, within the group of these patients under treatment, some may need more intensive treatment than others in order to arrive at the same health status score.

A quantification of this phenomenon is given in Table 5.4, which shows that there does not exist a monotonous relationship between PCS scores 2001 and observed costs 2002. For example, the mean PCS score of enrollees taking pharmaceutical drugs for HIV/Aids (PCG11) does not statistically significantly differ from the mean PCS score of those taking pharmaceutical drugs for asthma or COPD (PCG01) (two-sided t-test of equal means, p=0.278). Based on the PCS score alone, costs for enrollees with HIV/Aids or asthma/COPD are expected to be close to each other. However, observed health care costs of HIV/Aids patients are more than four times larger on average (two-sided t-test of equal means, p=0.000).

5.2 Estimation of the normative equation

The last column of Table 5.5 shows the coefficients after estimation of the normative equation (2.3), where observed costs are regressed on the S-type adjusters presented in Table 5.1.[95] The second column in Table 5.5 lists the estimated coefficients in case that the PCGs and DCGs are not included in the normative regression equation (2.3). The treatment effect as discussed in Section 5.1 is illustrated by the increase in R-squared from 7.62% to 19.58% that is caused by the inclusion of the PCGs and DCGs in the normative equation. Variation in expenditures caused by severity differences within treated groups of enrollees might be captured by estimating the coefficients of the SF-36 coefficients separately for each included disease and disorder (Hornbrook and Goodman 1996), but because of data limitations such interactions are not included in the normative equation in this study.[96]

It appears that most of the estimated coefficients corresponding to the SF-36 health status scales, the OECD limitations and in particular the self-reported chronic conditions are reduced after inclusion of the PCGs and DCGs. The estimated MH coefficient does not have the expected sign, whether the PCGs and

95. An insured can belong to multiple PCGs and/or multiple DCGs in order to better capture cost variation resulting from existing co-morbidities. Co-morbidities are also captured by the number of self-reported chronic conditions.

96. R^2 equals 7.03% if in addition to the PCGs and DCGs also age and sex are excluded from the normative equation. On the other hand, a normative equation with only age, sex, PCGs and DCGs included as predictors gives an R^2 equal to 17.89%.

Table 5.5: Estimated regression coefficients for the normative regression, exclusive and inclusive of age and gender, (multiple) PCGs and (multiple) DCGs.

	Estimated coefficients	
Explanatory variables	**PCGs and DCGS excluded**	**All relevant variables are included**
PF (Physical Functioning)	-22 *	-17 *
RP (Role Physical)	-4 *	-2
BP (Bodily Pain)	1	-3
GH (General Health)	-19 *	-9 *
VT (Vitality)	2	4
SF (Social Functioning)	-8 *	-6 *
RE (Role Emotional)	0	0
MH (Mental Health)	16 *	9 *
One self-reported OECD limitation	378 *	394 *
Two self-reported OECD limitations	647 *	657 *
Three or more self-reported OECD limitations	1221 *	929 *
One self-reported chronic condition	1720 *	916
Two self-reported chronic conditions	2642 *	1209 **
Three or more self-reported chronic conditions	4999 *	2450 *
M 15-24	---	---
M 25-34	-155	-196
M 35-44	-393	-438
M 45-54	442	371
M 55-64	389	210
M 65-74	1465 *	1038 *
M 75-84	2231 *	2015 *
M >=85	538	587
F 15-24	-147	-184
F 25-34	270	302
F 35-44	-111	-90
F 45-54	-87	-102
F 55-64	-4	-27
F 65-74	427	371
F 75-84	1048 *	1149 *
F >=85	-345	46
PCG01	---	1645
PCG02	---	1430
PCG03	---	282
PCG04	---	989
PCG05	---	2998
PCG06	---	2036

Explanatory variables	Estimated coefficients	
	PCGs and DCGS excluded	All relevant variables are included
PCG07	---	1898
PCG08	---	6699
PCG09	---	3086
PCG10	---	6705
PCG11	---	11783
PCG12	---	19574
DCG01	---	686
DCG02	---	4389
DCG03	---	3370
DCG04	---	4408
DCG05	---	2271
DCG06	---	6055
DCG07	---	3540
DCG08	---	6802
DCG09	---	5289
DCG10	---	14858
DCG11	---	6969
DCG12	---	8748
DCG13	---	74838
Intercept	3871 *	2967 *
R^2_{ADJ}	7.62 %	19.58 %

Note: PCG01=Asthma/COPD, PCG02=Epilepsy, PCG03=Crohn/Colitis Ulcerosa, PCG04=Cardiac disease, PCG05=Rheumatism, PCG06=Parkinson, PCG07=Diabetes (Type I), PCG08=Transplantation, PCG09=Cystic fibrosis, PCG10=Neuromuscular disorder, PCG11=HIV/Aids, PCG12=Renal disease/ESRD. See Appendix A for a description of the DCG classification used in this study. An insured can belong to multiple PCGs and/or multiple DCGs.
*,** The estimated coefficient is statistically significant from zero (two-sided t-test, * p <= 0.05, ** p <= 0.10).

DCGs are included or not, but remember that the presented weights reflect partial effects on health care costs such that part of the expected effect associated with mental health may already have been captured by other variables included in the regression. Moreover, if the PCS and MCS are included instead of the eight SF-36 subscales, the estimated coefficients with respect to PCS and MCS both have the appropriate negative sign.[97]

97. In order to capture the maximum extent of systematic variation possible, the SF-36 subscales PF, RP, BP, GH, VT, SF, RE and MH are included in the normative regression (2.3) instead of the summary scales PCS and MCS. See also Section 2.3.

For each SF-36 health status scale, the scale values are heavily skewed to the right (results not presented). Therefore transformations of the dummy variables for the self-reported health status scales were also tested as candidates to include in the normative equation. For each scale, dummy variables were created for the first, second and third quartile of the continuous metric scale values. Also, an alternative variant was tested where these dummy variables are interacted with their corresponding continuous metric scale values. For the dummy variables variant, only 6 out of 24 variables appeared statistically significant different from zero when included in the normative equation (two-sided t-test, $p <= 0.05$). For the combined dummy variables-continuous metric scale variant this was the case for only 4 out of 24 variables. Furthermore, for both transformations, an F-test of equality between the coefficients corresponding to the three quartile variables could not be rejected for 4 out of 8 of the scales ($p <= 0.05$) and explained variance appeared to be smaller than the 19.58% reported in Table 5.5. Therefore, the untransformed SF-36 health status scales were preferred as S-type adjusters to include in the normative equation.[98]

Given the estimated coefficients shown in Table 5.5, the normative costs can be derived following equation (2.4). In Table 5.6 normative costs 2002 are compared to observed costs 2002 for subgroups that are based on the S-type adjusters. Remember from the discussion in Section 2.3 that average normative costs are identical to average observed costs for these tabulated subgroups, provided the same subgroups of insured people are used in the tabulation as in the normative equation (2.4). This property of the ordinary least-squares technique holds exactly for each of the subgroups defined by the S-type adjusters as dummy variables. Tabulation of the subgroups defined by the number of self-reported chronic conditions, age and gender, PCGs and DCGs therefore show equality between normative costs and observed costs on average. Surprisingly, such equality does not hold for the subgroups defined by the number of self-reported OECD limitations, but this is because the number of self-reported OECD limitations had to be imputed for 1.8% of the total population who did not report it themselves.[99]

For continuous metric S-type adjusters this property of the ordinary least-squares technique does not hold for subgroups of survey respondents but only at the level of the total group of survey respondents. In order to compare normative costs to observed costs for subgroups in this case, these subgroups are constructed based

98. Conclusions in subsequent chapters do not change if these transformations are included in the normative equation nonetheless.

99. See also footnote 1.

Table 5.6: Normative costs compared to observed costs 2002, for subgroups of survey respondents based on the S-type adjusters from the normative equation (2.3)

Subgroups	Size of subgroup	Normative costs (pipy)	Observed costs (pipy)	Normative – observed costs	Normative / observed costs
Q1 PF scores	25.0%	3742	3650	92	1.025
Q1 RP scores	25.0%	3500	3479	21	1.006
Q1 BP scores	25.0%	3138	3206	-68	0.979
Q1 GH scores	25.0%	3512	3587	-76	0.979
Q1 VT scores	25.0%	2855	2905	-50	0.983
Q1 SF scores	25.0%	2953	3011	-58	0.981
Q1 RE scores	25.0%	2687	2702	-15	0.995
Q1 MH scores	25.0%	2330	2378	-48	0.980
Q2 PF scores	25.0%	1702	1649	53	1.032
Q2 RP scores	25.0%	1537	1583	-46	0.971
Q2 BP scores	25.0%	1680	1558	122	1.078
Q2 GH scores	25.0%	1662	1478	184 *	1.124 *
Q2 VT scores	25.0%	1747	1740	7	1.004
Q2 SF scores	25.0%	1835	1691	144 **	1.085 **
Q2 RE scores	25.0%	1471	1475	-4	0.998
Q2 MH scores	25.0%	1826	1771	55	1.031
Q3 PF scores	25.0%	948	1010	-61	0.939
Q3 RP scores	25.0%	1056	1060	-4	0.996
Q3 BP scores	25.0%	1141	1163	-22	0.981
Q3 GH scores	25.0%	1123	1098	25	1.023
Q3 VT scores	25.0%	1296	1244	52	1.041
Q3 SF scores	25.0%	1189	1280	-92	0.928
Q3 RE scores	25.0%	1589	1627	-38	0.977
Q3 MH scores	25.0%	1441	1386	54	1.039
Q4 PF scores	25.0%	621	705	-84 **	0.881 **
Q4 RP scores	25.0%	921	892	29	1.032
Q4 BP scores	25.0%	1055	1087	-32	0.970
Q4 GH scores	25.0%	718	851	-133 **	0.844 **
Q4 VT scores	25.0%	1116	1125	-8	0.993
Q4 SF scores	25.0%	1037	1032	5	1.005
Q4 RE scores	25.0%	1267	1210	57	1.047
Q4 MH scores	25.0%	1417	1479	-61	0.959
Number of self-reported OECD limitations					
0	77.8%	1218	1214	4	1.003
1	10.3%	2911	2904	7	1.002
2	4.9%	3954	3944	11	1.003
3+	5.2%	5277	5262	15	1.003

Subgroups	Size of subgroup	Normative costs (pipy)	Observed costs (pipy)	Normative – observed costs	Normative / observed costs
Imputed	1.8%	2038	2322	-285	0.877
Number of self-reported chronic conditions					
0	90.4%	1414	1414	0	1.000
1	8.1%	4592	4592	0	1.000
2	1.3%	6535	6535	0	1.000
3+	0.2%	9181	9181	0	1.000
M 15-24	4.2%	818	818	0	1.000
M 25-34	7.1%	691	691	0	1.000
M 35-44	6.9%	761	761	0	1.000
M 45-54	6.3%	1901	1901	0	1.000
M 55-64	5.7%	2378	2378	0	1.000
M 65-74	5.0%	3785	3785	0	1.000
M 75-84	2.3%	5045	5045	0	1.000
M >=85	0.3%	3742	3742	0	1.000
F 15-24	6.2%	812	812	0	1.000
F 25-34	11.3%	1256	1256	0	1.000
F 35-44	12.8%	1010	1010	0	1.000
F 45-54	11.3%	1450	1450	0	1.000
F 55-64	8.9%	1867	1867	0	1.000
F 65-74	7.2%	2814	2814	0	1.000
F 75-84	4.0%	4262	4262	0	1.000
F >=85	0.7%	3629	3629	0	1.000
PCG01	4.0%	5075	5075	0	1.000
PCG02	0.5%	4478	4478	0	1.000
PCG03	0.2%	3711	3711	0	1.000
PCG04	3.1%	6587	6587	0	1.000
PCG05	0.3%	5969	5969	0	1.000
PCG06	0.1%	6451	6451	0	1.000
PCG07	1.2%	6217	6217	0	1.000
PCG08	0.1%	11945	11945	0	1.000
PCG09	0.0%	7485	7485	0	1.000
PCG10	0.1%	11747	11747	0	1.000
PCG11	0.1%	13283	13283	0	1.000
PCG12	0.0%	40141	40141	0	1.000
DCG01	0.5%	4958	4958	0	1.000
DCG02	0.6%	9168	9168	0	1.000
DCG03	0.5%	8681	8681	0	1.000
DCG04	0.5%	11073	11073	0	1.000

Subgroups	Size of subgroup	Normative costs (pipy)	Observed costs (pipy)	Normative – observed costs	Normative / observed costs
DCG05	0.3%	8293	8293	0	1.000
DCG06	0.1%	11161	11161	0	1.000
DCG07	0.3%	10540	10540	0	1.000
DCG08	0.2%	13437	13437	0	1.000
DCG09	0.0%	11058	11058	0	1.000
DCG10	0.1%	23637	23637	0	1.000
DCG11	0.1%	15094	15094	0	1.000
DCG12	0.1%	23892	23892	0	1.000
DCG13	0.0%	84125	84125	0	1.000
Total	18617	1753	1753	0	1.000

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

on the quartiles of the continuous metric S-type adjuster. This leads to a set of four subgroups for each of the eight continuous metric SF-36 scales, as presented in Table 5.6. It may be expected that normative costs will be largest for those with the worst health status, i.e. for those insured assigned to the subgroup associated with the bottom or first quartile of the SF-36 scale scores.

For the insured who belong to the bottom 25% or to the bottom 50% of the eight SF-36 health status scales, Table 5.6 shows that on average normative costs are not statistically significantly different from observed costs. Note that this result does not follow from the ordinary least squares technique merely by construction, a property which only holds for those insured with average SF-36 scale scores. Given an estimation of normative costs following equation (2.4), the next step is to find an answer to the question to what extent normative costs are captured if only a limited set of measures of the S-type risk factors is feasible to include as REF adjusters in the REF equation. This is usually the case with the conventional risk equalization approach. An answer to this first research question in the context of the 2004 Dutch REF equation is given in Chapter Six.

5.3 Conclusions

In this study, it is assumed that the sponsor desires cross-subsidization among subgroups defined by the S-type risk factors age, gender and health status. Administrative data on age and gender are used as measures of the S-type risk factors age and gender. The eight SF-36 scales, the number of self-reported OECD

limitations and the number of self-reported chronic conditions from the Agis Health Survey 2001 are used as measures of the S-type risk factor health status.

In addition, (multiple) PCGs and (multiple) DCGs are included in the normative equation (2.3) in order to capture cost variation between disease groups that may not be captured by the other health status measures. From the normative regression results it appears that adding the PCGs and DCGs improves the adjusted R^2 from 7.62% to 19.58%. It is important to stress here that the S-type adjusters are included as measures of the S-type risk factors in the normative regression (2.3) for normative reasons only and not because of their additional explanatory power per se.

In this chapter, the chosen S-type adjusters are first described statistically and cross-tabulated in order to check whether the theoretical motivation to include them in the normative equation also holds for the present study sample. This appears to be the case indeed. After this validity check, the normative equation (2.3) is estimated by ordinary least squares and normative costs are calculated following equation (2.4).

Risk-adjusted premium subsidies must be based on normative costs in order to satisfy the criterion of the effectiveness of the risk-adjusted premium subsidies. The difference between normative costs and observed costs is zero for the subgroups defined by the S-type adjusters by construction. This identity holds because the coefficients of equation (2.3) are estimated by ordinary least-squares.

The normative costs derived in this chapter form the basis for the analysis in Chapters Six and Seven. In Chapter Six, REF predicted costs following from REF equation (2.2) are tested against normative costs as derived above. In Chapter Seven, alternative specifications of the REF model are tested against normative costs as an illustration of the different ways to improve upon the results of the conventional specification of the 2004 Dutch REF equation.

Appendix A5.1: Diagnosic Cost Group classification

Table A5.1: Diagnostic Cost Group (DCG) classification by diagnosis group (DxG) [a,b]

Dxg	Description	DCG
175	Haemodialysis at home	13
163	Major Organ Transplant Status	12
176	Artificial respiration at home	12
115	Renal Failure/Nephritis	12
111	Pulmonary Fibrosis and Bronchiectasis	12
144	Spinal Cord Injury	12
15	Blood, Lymphatic Cancers/Neoplasms	12
9	Liver/Pancreas/Esophagus Cancer	11
7	Metastatic Cancer	11
33	End Stage Liver Disorders	11
20	Brain/Nervous System Cancer	11
27	Diabetes with Chronic Complications	10
13	Lung Cancer	10
134	Decubitus and Chronic Skin Ulcers	9
55	Blood/Immune Disorders	9
8	Mouth/Pharynx/Larynx/Other Respiratory Cancer	9
3	HIV/AIDS	8
105	Chronic Obstructive Pulmonary Disease	8
10	Stomach, Small Bowel, Other Digestive Cancer	8
72	Paralytic and Other Neurologic Disorders	7
34	Cirrhosis, Other Liver Disorders	7
48	Rheumatoid Arthritis and Connective Tissue Disease	7
70	Degenerative Neurologic Disorders	7
12	Rectal Cancer	7
87	Paroxysmal Ventricular Tachycardia	7
173	Chemotherapy	7
17	Cancer of Placenta/Ovary/Uterine Adnexa	7
21	Other Cancers	7
95	Atherosclerosis of Major Vessel	6
76	Coma and Enccphalopathy	6
98	Peripheral Vascular Disease	6
19	Cancer of Bladder, Kidney, Urinary Organs	6
158	Artificial Opening of Gastrointestinal Tract Status	6
89	Congestive Heart Failure	5
32	Pancreatitis/Other Pancreatic Disorders	5
77	Valvular and Rheumatic Heart Disease	5
91	Cerebral Hemorrhage	5
22	Benign Brain/Nervous System Neoplasm	5
37	Stomach Disorders	5
41	Inflammatory Bowel Disease	4

Dxg	Description	DCG
79	Hypertension, Complicated	4
49	Bone/Joint Infections/Necrosis	4
96	Aortic and Other Arterial Aneurysm	4
171	Major Congenital Disorders	4
174	Radiation therapy	4
18	Cancer of Prostate/Testis/Male Genital Organs	4
93	Stroke	4
11	Colon Cancer	4
92	Precerebral Arterial Occlusion	4
25	Diabetes with No or Unspecified Complications	3
88	Cardio-Respiratory Failure and Shock	3
83	Unstable Angina	3
35	Diseases of Esophagus	3
73	Epilepsy and Other Seizure Disorders	3
43	Diverticula of Intestine	3
14	Breast Cancer	3
97	Thromboembolic Vascular Disease	3
110	Asthma	2
84	Angina Pectoris	2
80	Coronary Atherosclerosis	2
16	Cancer of Uterus/Cervix/Female Genital Organs	2
81	Post-Myocardial Infarction	2
153	Brain Injury	2
36	Peptic Ulcer	2
26	Diabetes with Acute Complications/Hypoglycemic	1
86	Atrial Arrhythmia	1
85	Heart Rhythm and Conduction Disorders	1
50	Osteoarthritis	1
150	Internal Injuries/Traumatic Amputations/Third Degree	1

Source: Van Vliet and Prinsze (2003, Part VII, Table 4)

[a] There are no claims associated with DxG 166 and DxG 167 present in the research sample used for this study, and DxG 81 (acute myocardial infarction) and DxG 82 (post-myocardial infarction) are combined into one DxG as they can not be identified separately from the hospital claims data. In total, there are 173 DxGs (incl. DxG 173, 174, 175 and 176 which are hospital procedures that are relevant in the Dutch setting).

[b] See Van Vliet and Prinsze (2003, Part VII, Appendix A) for the classification of ICD-codes in diagnosis groups (DxG), originating from Pope et al. (1999).

Chapter 6

Testing the REF equation for effectiveness

In Section 6.1, REF predicted costs are tested against normative costs for the subgroups defined by the S-type adjusters in order to determine the extent to which the REF adjusters included in the 2004 Dutch REF equation satisfy the criterion of effectiveness. In Section 6.2 an empirical application of the omitted variables approach to adjust the REF weights is presented, such that the gap between REF predicted costs and normative costs is reduced for the subgroups defined by the REF adjusters. In Section 6.3 a normative adjustment procedure is applied that completely removes the gap between REF predicted costs and normative costs for the subgroups defined by the REF adjusters.

6.1 A COMPARISON OF REF PREDICTED COSTS AND NORMATIVE COSTS

In Table 6.1 the REF adjusters are presented that are also present in the 2004 Dutch REF equation. In this study, a smaller number of categories is employed than in the 2004 Dutch REF equation with respect to age and eligibility: eight instead of nineteen age categories are applied for both men and women, and no interactions between eligibility and age are included. This is done because of limitations of the sample size. See Chapter Five for a motivation of the choice of the number of age categories.

Table 6.1: REF adjusters 2001 included in equation (2.1)

REF adjusters	Number of 0/1 dummies	Reference category
Gender*age	2 * 8 - 1	Male enrollees, 15-24 years of age
Eligibility	5 - 1	Employed enrollees
Region	10 - 1	Regional cluster 10
PCGs (single)	12	No PCG
DCGs (single)	13	No DCG

To calculate the REF predicted costs, equation (2.1) must first be estimated with the full set of adjusters shown in Table 6.1. The estimation results are shown in Table 6.2. The adjusted R^2 for this estimated regression equation equals 17.93%, which is in line with results on the Dutch REF equation reported elsewhere (see e.g. Lamers, and Van Vliet 2004).[100]

100. In this study, maximization of the adjusted R^2 is not an inherent goal. The REF weights must be determined such that REF predicted costs are as close as possible to normative costs (instead of observed costs).

Table 6.2: Estimation results for the REF equation (2.2).

REF adjusters	REF weights
Intercept	623
M 15-24 (reference category)	---
M 25-34	-209
M 35-44	-316
M 45-54	464
M 55-64	380
M 65-74	1493
M 75-84	2796
M >=85	1598
F 15-24	-109
F 25-34	345
F 35-44	0
F 45-54	194
F 55-64	274
F 65-74	1003
F 75-84	2289
F >=85	1662
Disabled	1437
Employed (reference category)	---
Social welfare	211
Unemployed	214
Retired	341
Self-Employed	-197
Region 1	457
Region 2	262
Region 3	138
Region 4	239
Region 5	37
Region 6	-121
Region 7	-31
Region 8	-164
Region 9	29
Region 10 (reference category)	---
No PCG (reference category)	---
PCG01	1883
PCG02	1803
PCG03	1098
PCG04	2001
PCG05	3848

REF adjusters	REF weights
PCG06	3199
PCG07	3366
PCG08	7791
PCG09	3823
PCG10	8030
PCG11	11895
PCG12	20748
No DCG (reference category)	---
DCG01	1356
DCG02	6319
DCG03	3565
DCG04	5591
DCG05	4262
DCG06	7820
DCG07	6038
DCG08	8869
DCG09	7983
DCG10	18152
DCG11	12626
DCG12	9050
DCG13	77982
R^2_{ADJ}	17.93%

Note: PCG01=Asthma/COPD, PCG02=Epilepsy, PCG03=Crohn/Colitis Ulcerosa, PCG04=Cardiac disease, PCG05=Rheumatism, PCG06=Parkinson, PCG07=Diabetes (Type I), PCG08=Transplantation, PCG09=Cystic fibrosis, PCG10=Neuromuscular disorder, PCG11=HIV/Aids, PCG12=Renal disease/ESRD. An insured can belong to a single PCG and/or a single DCG only.

6.1.1 Comparing subgroups defined by the S-type adjusters

Given the REF weights from Table 6.2, REF predicted costs of an enrollee are calculated following equation (2.2). These REF predicted costs can then be compared with normative costs as derived in Chapter Five.[101] The ultimate test procedure to be followed in this case is to make such a comparison for the subgroups of insured based on the S-type adjusters that are included in the normative equation (2.3). If equality holds for all such subgroups, the Dutch REF equation fully satisfies the

101. The usual approach to test REF equations in the literature is to compare REF predicted costs with observed costs instead of normative costs. The main relevance of that approach is that under premium regulation tabulated predictable losses and profits identify incentives for risk selection (and their adverse effects) for subgroups of enrollees. Note that these tabulated predictable profits and losses may be induced both by S-type and N-type risk factors. In this study, it is assumed that premiums are not regulated.

Table 6.3: REF predicted costs compared to normative costs 2002, for subgroups of survey respondents defined by S-type adjusters from the normative equation (2.3): age and gender

Subgroups of enrollees	Size of subgroup	REF predicted costs (pipy)	Normative costs (pipy)	REF predicted – normative costs	REF predicted / normative costs
M 15-24	4.2%	818	818	0	1.000
M 25-34	7.1%	691	691	0	1.000
M 35-44	6.9%	761	761	0	1.000
M 45-54	6.3%	1901	1901	0	1.000
M 55-64	5.7%	2378	2378	0	1.000
M 65-74	5.0%	3785	3785	0	1.000
M 75-84	2.3%	5045	5045	0	1.000
M >=85	0.3%	3742	3742	0	1.000
F 15-24	6.2%	812	812	0	1.000
F 25-34	11.3%	1256	1256	0	1.000
F 35-44	12.8%	1010	1010	0	1.000
F 45-54	11.3%	1450	1450	0	1.000
F 55-64	8.9%	1867	1867	0	1.000
F 65-74	7.2%	2814	2814	0	1.000
F 75-84	4.0%	4262	4262	0	1.000
F >=85	0.7%	3629	3629	0	1.000
Total	100.0%	1753	1753	0	1.000

Note: The age and gender categories applied in this table are used as 0/1 dummy explanatory variables in the estimation of both the REF equation (2.1) and the normative equation (2.3). Age categories below 15 years of age are not observed as the research sample only contains survey respondents of 16 years and older.

criterion of effectiveness. However, in this study this is ruled out by definition as the subgroups defined by the S-type adjusters differ from the subgroups defined by the REF adjusters.

In Table 6.3 REF predicted costs and normative costs are compared for each gender and age subgroup of enrollees. More specifically, the difference with and ratio to normative costs is tabulated.[102] The difference between REF predicted and normative costs appears to be equal to zero for all age-gender-subgroups of enrollees. The explanation is that age and gender are not only included in the REF equation but also in the normative equation, and in both cases the very same classification is applied as in Table 6.3. As a consequence, at the age and gender group level, both average REF predicted costs and normative costs are equal to

102. The ratio between REF predicted costs and normative costs must not be confused with the ratio of REF predicted costs and observed costs, which is the traditional definition of a predictive ratio in the risk adjustment literature. See also the previous footnote.

average observed costs and therefore also to each other. The same explanation holds for the result that the ratio of REF predicted and normative costs equals one for all tabulated age-gender-subgroups of enrollees. These results show that the normative test of effectiveness of the risk-adjusted premium subsidies at the age and gender subgroup level is passed by the conventional REF approach to risk adjustment by construction.

Pharmacy-based and diagnostic cost groups (i.e. PCGs and DCGs) are also included both in the REF equation and the normative equation, and therefore equality between average REF predicted and normative costs might be expected here too. However, some deviations of REF predicted costs from normative costs are observed in Tables 6.4 and 6.5. The explanation is that in order to estimate REF

Table 6.4: REF predicted costs compared to normative costs 2002, for subgroups of survey respondents defined by S-type adjusters from the normative equation (2.3): Pharmaceutical Cost Groups (PCGs) that are not rank-ordered

Subgroups of enrollees	Size of subgroup	REF predicted costs (pipy)	Normative costs (pipy)	REF predicted – normative costs	REF predicted / normative costs
No PCG	91.2%	1391	1382	9	1.006
PCG01	3.4%	4653	5075	-422 *	0.917 *
PCG02	0.5%	4414	4478	-65	0.986
PCG03	0.2%	4192	3711	482	1.130
PCG04	2.8%	6326	6587	-261	0.960
PCG05	0.3%	5957	5969	-12	0.998
PCG06	0.1%	6411	6451	-40	0.994
PCG07	1.2%	6169	6217	-48	0.992
PCG08	0.1%	11738	11945	-208	0.983
PCG09	0.0%	7485	7485	0	1.000
PCG10	0.1%	11747	11747	0	1.000
PCG11	0.1%	13283	13283	0	1.000
PCG12	0.0%	40141	40141	0	1.000
Total	100.0%	1753	1753	0	1.000

Note: PCG01=Asthma/COPD, PCG02=Epilepsy, PCG03=Crohn/Colitis Ulcerosa, PCG04=Cardiac disease, PCG05=Rheumatism, PCG06=Parkinson, PCG07=Diabetes (Type I), PCG08=Transplantation, PCG09=Cystic fibrosis, PCG10=Neuromuscular disorder, PCG11=HIV/Aids, PCG12=Renal disease/ESRD. The PCGs applied in this table are used as 0/1 dummy explanatory variables in the estimation of both the REF equation (2.1) and the normative equation (2.3). In the former equation they are rank-ordered, in the latter equation they are not.

* Difference between average predicted and normative costs is statistically significant (two-sided t-test, $p <= 0.05$).

Table 6.5: REF predicted costs compared to normative costs 2002, for subgroups of survey respondents defined by S-type adjusters from the normative equation (2.3): Diagnostic Cost Groups (DCGs) that are not rank-ordered

Subgroups of enrollees	Size of subgroup	REF predicted costs (pipy)	Normative costs (pipy)	REF predicted – normative costs	REF predicted / normative costs
No DCG	97.2%	1522	1522	0	1.000
DCG01	0.4%	4774	4958	-185	0.963
DCG02	0.5%	9022	9168	-147	0.984
DCG03	0.4%	7199	8681	-1482 *	0.829 *
DCG04	0.4%	9964	11073	-1109	0.900
DCG05	0.3%	8554	8293	261	1.031
DCG06	0.1%	11063	11161	-98	0.991
DCG07	0.2%	10618	10540	78	1.007
DCG08	0.2%	13560	13437	123	1.009
DCG09	0.0%	11058	11058	0	1.000
DCG10	0.1%	23347	23637	-290	0.988
DCG11	0.1%	15094	15094	0	1.000
DCG12	0.1%	21175	23892	-2717	0.886
DCG13	0.0%	84125	84125	0	1.000
Total	100.0%	1753	1753	0	1.000

Note: The DCGs applied in this table are used as 0/1 dummy explanatory variables in the estimation of both the REF equation (2.1) and the normative equation (2.3). In the former equation they are rank-ordered, in the latter equation they are not.

* Difference between average predicted and normative costs is statistically significant (two-sided t-test, $p <= 0.05$).

predicted costs a different classification of PCGs and DCGs is applied than when estimating normative costs. More specifically, an insured can belong to a single PCG and/or a single DCG in the context of the REF equation, whereas in the context of the normative equation an insured can belong to multiple PCGs and/or multiple DCGs. The restriction with respect to the REF equation is applied in order to mitigate the incentives for strategic upcoding behavior by providers in case of existing comorbidities.[103]

Tables 6.4 and 6.5 show the differences between REF predicted costs and normative costs, for subgroups of insured people classified by PCGs and DCGs which are not rank-ordered. These tables therefore show the extent to which REF predicted costs deviate from normative costs for these subgroups as a direct consequence of the application of rank-ordering to PCGs and DCGs for implementation purposes.

103. See Section 5.1 for a description of the rank-ordering procedure applied in this study. See also Lamers and Van Vliet (2003) for a description of this iterative procedure in the context of the construction of the pharmacy-based cost groups (PCGs).

Tables 6.4 and 6.5 show that REF predicted costs for enrollees assigned to PCG01 are 422 euro (8.3%) below normative costs, for enrollees assigned to DCG03 they are 1482 euro (17.1%) short of normative costs. For the other PCG and DCG subgroups the difference between REF predicted costs and normative costs is not statistically significant from zero (two-sided t-test, $p>0.05$). This also holds for enrollees not assigned to any PCG and those not assigned to any DCG. These results show that the normative test of effectiveness of the risk-adjusted premium subsidies is passed for almost all PCG and DCG subgroups.

Table 6.6 is based on subgroups that are defined by the quartiles of the eight SF-36 scale scores, the number of self-reported OECD limitations, and the number of self-reported chronic conditions. It appears that REF predicted costs fall short of normative costs for those insured people with first quartile SF-36 scale scores and those who reported the existence of one or more OECD limitations or chronic conditions. In other words, those with a relatively bad health status are

Table 6.6: REF predicted costs compared to normative costs 2002, for subgroups of survey respondents defined by S-type adjusters from the normative equation (2.3): SF-36 scores, number of self-reported OECD limitations and number of self-reported chronic conditions [a]

Subgroups of enrollees	Size of subgroup	REF predicted costs (pipy)	Normative costs (pipy)	REF predicted – normative costs	REF predicted / normative costs
Q1 PF scores	25.0%	2990	3742	-752 *	0.799 *
Q1 RP scores	25.0%	2879	3500	-621 *	0.823 *
Q1 BP scores	25.0%	2510	3138	-628 *	0.800 *
Q1 GH scores	25.0%	2912	3512	-599 *	0.829 *
Q1 VT scores	25.0%	2436	2855	-419 *	0.853 *
Q1 SF scores	25.0%	2439	2953	-514 *	0.826 *
Q1 RE scores	25.0%	2370	2687	-318 *	0.882 *
Q1 MH scores	25.0%	2109	2330	-221 *	0.905 *
Q2 PF scores	25.0%	1826	1702	124 *	1.073 *
Q2 RP scores	25.0%	1613	1537	77 *	1.050 *
Q2 BP scores	25.0%	1720	1680	40	1.024
Q2 GH scores	25.0%	1708	1662	46 **	1.028 **
Q2 VT scores	25.0%	1744	1747	-3	0.999
Q2 SF scores	25.0%	1835	1835	1	1.000
Q2 RE scores	25.0%	1574	1471	103 *	1.070 *
Q2 MH scores	25.0%	1830	1826	4	1.002
Q3 PF scores	25.0%	1241	948	292 *	1.308 *
Q3 RP scores	25.0%	1344	1056	288 *	1.272 *
Q3 BP scores	25.0%	1393	1141	252 *	1.221 *

Subgroups of enrollees	Size of subgroup	REF predicted costs (pipy)	Normative costs (pipy)	REF predicted – normative costs	REF predicted / normative costs
Q3 GH scores	25.0%	1348	1123	225 *	1.200 *
Q3 VT scores	25.0%	1459	1296	163 *	1.126 *
Q3 SF scores	25.0%	1439	1189	250 *	1.211 *
Q3 RE scores	25.0%	1707	1589	118 *	1.074 *
Q3 MH scores	25.0%	1534	1441	93 *	1.065 *
Q4 PF scores	25.0%	957	621	336 *	1.541 *
Q4 RP scores	25.0%	1178	921	256 *	1.278 *
Q4 BP scores	25.0%	1391	1055	336 *	1.319 *
Q4 GH scores	25.0%	1046	718	328 *	1.457 *
Q4 VT scores	25.0%	1375	1116	259 *	1.232 *
Q4 SF scores	25.0%	1300	1037	263 *	1.254 *
Q4 RE scores	25.0%	1363	1267	97 *	1.076 *
Q4 MH scores	25.0%	1540	1417	123 *	1.087 *
Number of self-reported OECD limitations [b]					
0	77.8%	1422	1218	204 *	1.168 *
1	10.3%	2541	2911	-370 *	0.873 *
2	4.9%	3038	3954	-916 *	0.768 *
3+	5.2%	3832	5277	-1445 *	0.726 *
Imputed	1.8%	2025	2038	-13	0.994
Number of self-reported chronic conditions [b]					
0	90.4%	1508	1414	93 *	1.066 *
1	8.1%	3814	4592	-778 *	0.831 *
2	1.3%	5267	6535	-1268 *	0.806 *
3+	0.2%	6670	9181	-2511 *	0.726 *
Total	100.0%	1753	1753	0	1.000

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

[a] The weighted average of the absolute values of the differences between REF predicted costs and normative costs for the tabulated subgroups equals 251. If the subgroups defined by the S-type adjusters age, gender, PCGs and DCGs are also taken into account (see the Tables 6.3, 6.4 and 6.5) then this figure equals 198.

[b] The sizes of these subgroups sum up to 100%.

undercompensated if the risk-adjusted premium subsidies are based on the REF adjusters. On the other hand, risk-adjusted premium subsidies will overcompensate most other tabulated subgroups of insured people if based on REF predicted costs. It must be concluded that the 2004 Dutch REF equation needs to be improved in order to safeguard affordability for high-risk individuals.

In order to determine the performance of the 2004 Dutch REF equation, first the statistic given by equation (2.6) should be calculated for all subgroups defined by the S-type adjusters. According to Table 6.6, footnote a, a weighted average of (the absolute values of) these subgroup statistics is equal to 198. This figure can be compared to the figure of 687 that would result in the absence of any cross-subsidies between the healthy and the sick insured people.[104] The risk-adjusted premium subsidies induced by the 2004 Dutch REF equation can therefore be estimated to be in line with the policy goals of the Dutch government up to an extent of (1-198/687) x 100% = 71.2%.

6.1.2 Comparing subgroups defined by the REF adjusters

In practice, the risk-adjusted premium subsidies are based on the subgroups defined by the feasible set of REF adjusters. Ideally, for each subgroup defined by a REF adjuster, REF predicted costs coincides with normative costs on average. This equality holds for the subgroups age and gender by construction, as these are included as adjusters in both the REF equation (2.2) and the normative equation (2.4).

The REF adjusters eligibility and region are included in the 2004 Dutch REF equation under the assumption that these variables can be seen as proxies for health status differences for which the Dutch government desires cross-subsidization. For example, it is expected that disabled enrollees have to cope with worse health conditions than enrollees being self-employed. Higher REF predicted costs for disabled enrollees and lower predicted costs for self-employed enrollees are then attributed to these differences in health conditions and compensated for by including eligibility as a adjuster in the REF equation.

However, with respect to the subgroups defined by eligibility and region, it is still an open question whether these cost differences must be entirely or only partially attributed to S-type risk factors. To the extent that cost differences can be attributed to N-type risk factors, the current REF weights are incorrect and induce undesired risk-adjusted premium subsidies to Dutch insured people. An answer to this empirical question can be given by comparing REF predicted costs with normative costs. Deviations from normative costs within a subgroup may then be attributed to cost variation caused by the N-type risk factors. Ideally, such deviations are avoided.[105]

---

104. The figure of 687 can be found in Table A6.1, footnote a.

105. Note that the REF adjusters eligibility and region are not included in the normative equation (2.3), therefore REF predicted costs are not equal to normative costs for these subgroups merely by construction. This is also the case with respect to subgroups defined by the REF adjusters PCGs

Table 6.7 shows that average REF predicted costs differ from average normative costs for most tabulated subgroups. REF predicted costs for disabled enrollees are 3204 euro on average, i.e. 420 euro (15%) above average normative costs of 2783 euro. This means that, conditional on the specific composition of the subgroup of disabled enrollees in the research sample, risk-adjusted premium subsidies should be based on 2783 instead of 3204 euro in order to induce the risk-adjusted premium subsidies that Dutch government desires. The overcompensation of disabled enrollees by 420 euro must be attributed to N-type risk factors.

Table 6.7 also shows that REF predicted costs for enrollees on social welfare and for self-employed enrollees are both 16% below normative costs, which amount to 2015 euro and 1001 euro respectively. This discrepancy means that these subgroups are undercompensated if risk-adjusted premium subsidies are based on REF predicted costs. Furthermore, before 2004, the REF adjuster insurance eligibility was defined such that employed and self-employed insured people belonged to the same subgroup in the Dutch REF equation. Much political discussion existed about the hypothesis that S-type risk factors caused REF predicted costs for the subgroup of self-employed enrollees to be lower than those for the subgroup of employed enrollees. Due to a lack of a proper theoretical framework it was not possible to properly test this hypothesis and a political decision was made to include employed and self-employed enrollees as separate subgroups in the 2004 Dutch REF equation. This is only justified under the hypothesis that the observed cost differences between these two subgroups are entirely caused by

Table 6.7: REF predicted costs compared to normative costs 2002, for subgroups of survey respondents defined by the REF adjusters from the REF equation (2.1): insurance eligibility

Subgroups of enrollees	Size of subgroup	REF predicted costs (pipy)	Normative costs (pipy)	REF predicted – normative costs	REF predicted / normative costs
Disabled	9.0%	3204	2783	420 *	1.151 *
Employed	59.5%	965	989	-24 **	0.976 **
Social welfare	4.1%	1689	2015	-327 *	0.838 *
Unemployed	4.2%	1597	1703	-106	0.938
Retired	20.5%	3573	3579	-5	0.998
Self-Employed	2.8%	839	1001	-162 *	0.838 *
Total	100.0%	1753	1753	0	1.000

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

and DCGs, because these subgroups are included in a non-rankordered manner in the normative equation. On the other hand, equality does hold by construction for subgroups defined by the REF adjusters gender and age.

Table 6.8: REF predicted costs compared to normative costs 2002, for subgroups of survey respondents defined by the REF adjusters from the REF equation (2.1): ten regional clusters of 2001 ZIP codes

Subgroups of enrollees	Size of subgroup	REF predicted costs (pipy)	Normative costs (pipy)	REF predicted – normative costs	REF predicted / normative costs
Region 1	7.2%	2052	1807	245 *	1.136 *
Region 2	20.6%	1976	1893	84	1.044
Region 3	9.5%	1639	1594	44	1.028
Region 4	9.2%	1847	1758	89 **	1.051 **
Region 5	14.6%	1699	1738	-39	0.978
Region 6	9.9%	1582	1772	-190 *	0.893 *
Region 7	16.7%	1649	1742	-94 *	0.946 *
Region 8	2.8%	1335	1506	-171 *	0.886 *
Region 9	3.3%	1550	1630	-80	0.951
Region 10	6.1%	1682	1680	3	1.002
Total	100.0%	1753	1753	0	1.000

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

S-type risk factors. However, Table 6.7 now reveals that although REF predicted costs are lower for the latter subgroup, normative costs are not. Therefore, the aforementioned hypothesis must be refuted: in this study sample REF predicted costs are lower for the subgroup of self-employed enrollees than for the subgroup of employed enrollees because of N-type instead of S-type risk factors. The conclusion is that the REF adjuster insurance eligibility fails to adequately capture S-type cost variation for self-employed enrollees.[106]

In Table 6.8 REF predicted costs and normative costs are compared in order to determine whether the four-digit regional ZIP code classification applied in the 2004 Dutch REF equation is a valid measure of the S-type risk factors. It appears that REF predicted costs for enrollees living in the first regional cluster of ZIP codes are 13.6% above normative costs, whereas regional clusters 6, 7 and 8 contain enrollees for which REF predicted costs lie between 5.4% and 11.4% below normative costs on average. It is concluded that part of the cost variation captured by the regional REF adjuster must be attributed to N-type risk factors.

106. Note that this does not necessarily mean that the decision to define a separate subgroup for self-employed enrollees was wrong. In particular, it would come at the expense of REF predicted costs for the employed enrollees if it was decided otherwise. The best strategy to overcome the misalignment of REF predicted costs with normative costs for the self-employed is to apply a normatively adjusted REF weight instead of the original REF weight, a procedure which will be applied in Section 6.3.

An explanation for this phenomenon is that the regional REF adjuster in the 2004 Dutch REF equation is constructed such that it not only compensates for cost variation defined by the S-type adjuster health status, but also compensates for cost variation caused by some N-type risk factors for which it is assumed that in the short term these can (almost) not be influenced by insurers' policies (e.g. the density and prices of health care providers).[107] In other words, the regional REF adjuster in the Dutch REF equation is based on another categorization of S-type and N-type risk factors than the categorization applied in this study.

In conclusion, the 2004 Dutch REF equation needs to be improved in order to satisfy the criterion of effectiveness. This conclusion is based on a comparison of REF predicted costs and normative costs for the subgroups defined by the S-type adjusters. Furthermore, given that the risk-adjusted premium subsidies are based on the subgroups defined by the REF adjusters in practice, deviations of REF predicted from normative costs for these subgroups should be avoided. Otherwise, this may lead to undesirable compensation for cost variation caused by N-type risk factors. This appears to hold in particular for some of the subgroups defined by the REF adjusters eligibility and region. In Section 6.2 the estimated REF weights will be adjusted under an omitted variables approach to better align REF predicted costs with normative costs for the subgroups defined by these REF adjusters.

6.2 An adjustment of the REF weights by the omitted variables approach

In Section 6.1 it was shown that risk-adjusted premium subsidies for the subgroups defined by eligibility and region would induce N-type cost variation if based on REF predicted costs. Under their structural form interpretation of the REF equation, Schokkaert, Dhaene and Van de Voorde (1998) and Schokkaert and Van de Voorde (2000, 2004) advocate an omitted variable approach to adjust the REF weights in order to avoid this compensation for N-type cost variation. The omitted variables approach may be called useful if it appears that the gap between REF predicted costs and normative costs disappears if the REF weights adjusted for the N-type risk factors are applied instead of the original REF weights. The omitted variables approach may be recommended for application in practice, to the extent that the deviation from normative costs disappears.

---

107. See Section 3.3 for a complete list of the risk factors on which the regional REF adjuster is based.

In this study, the following three administrative variables are included to capture the effects of the N-type risk factors: hospital output prices, distance to the general practitioner and distance to the hospital. These N-type adjusters are all measured at the four-digit ZIP code level[108]:

- Hospital output prices 2002 are defined as a weighted average of hospital fees for a one day hospital stay per ZIP code, where the weights are the number of 2001 outpatient contacts that Agis enrollees living in that ZIP code had with hospitals.[109] Hospital output prices differ substantially between hospitals and may be seen as N-type adjusters that induce cost-variation for which insurers are held responsible.[110]
- The distance to the nearest health care facility may be an important determinant of health care use and measures health care accessibility and time price of health care use for an enrollee. Distances to the nearest hospital and GP are measured between the centroids of four-digit ZIP codes, and are equal to zero if the health care facility's ZIP code equals that of the enrollee. Note that in this study, distance to the nearest health care facility is treated as an N-type adjuster.[111]

It should be noted that the regional REF adjuster is based on hospital output prices, distance to the general practitioner and distance to the hospital by construction.[112] In other words, contrary to the assumption in this study, these are treated as measures of S-type risk factors in the 2004 Dutch model. Therefore, it is expected that application of the omitted variables approach will have the largest impact on the size of the regional REF weights.

---

108. The Dutch regional ZIP code classification contains six digits in total. Only the first four positions are used here.

109. In order to calculate these weights, not only outpatient contacts of Agis enrollees present in the research sample are used but those of all Agis enrollees present in the 2001 sickness fund population.

110. As an alternative approach, hospital costs – as part of the dependent variable in the REF equation - could have been recalculated by applying a uniform outpatient hospital tariff to all enrollees with hospital costs in the research sample instead of the hospital specific tariffs that are applied in practice. Furthermore, note that tariffs are largely uniform for non-hospital costs in the present dataset and therefore a similar procedure is not needed for that type of costs.

111. For 18 out of 18617 enrollees distances to the nearest GP and hospital are not available at a four-digit ZIP code level. In these cases, a weighted average of distances for those four-digit ZIP codes that start with the same three digits (14 cases) or two digits (4 cases) has been imputed.

112. See Section 3.3 for a description of the construction of the regional REF adjuster.

Table 6.9 shows the estimation results for the REF equation and the adjusted REF equation if N-type adjusters are included during the estimation phase. The unadjusted REF weights presented in Table 6.9 are copied from Table 6.2. Percentages of explained variance are identical, i.e. indifferent from the choice to include or exclude the chosen N-type adjusters during the estimation phase.

The differences between unadjusted REF weights and REF weights from a regression including N-type adjusters are less than three percent in case of the REF adjusters age, gender, and eligibility, with the exception of the categories females between 15 and 24 years of age (10.1%) and enrollees on welfare (-10.9%). The estimated coefficients of PCGs and DCGs even change less than one percent when including the administrative N-type adjusters.

As expected, REF weights with respect to the regional REF adjuster change quite substantially if the N-type adjusters are included, ranging from about half the unadjusted REF weight to more than five times.[113] These weights are reduced relatively most for the higher ranked regional clusters (see the column "Change in weights"). The explanation is that hospital output prices are relatively high and distances to health care facilities are relatively small for enrollees living in these higher ranked clusters. As output price and accessibility differences are compensated for via the regional REF adjuster in the 2004 Dutch model, treating them as N-type adjusters instead reverses these compensations.[114]

The estimated REF weights of the included N-type adjusters show that if hospital output price is 100 euro larger than in another ZIP code, then REF predicted costs are 33 euro larger and adjusted REF predicted costs are 7 euro larger, ceteris paribus. For each kilometer that an enrollee lives closer to a GP, REF predicted costs are 112 euro higher, ceteris paribus. For each kilometer living closer to a hospital, REF predicted costs are 5 euro higher, ceteris paribus. Note, once again, that the 2004 Dutch model compensates for these effects.

Table 6.10 presents the results for subgroups defined by the 2004 Dutch regional REF adjuster. It appears that the deviations from normative costs are reduced if the N-type adjusters are taken into account during the estimation phase, except

113. The fact that the regional classification is partly based on the distance variables may be a partial explanation for this. In the 2004 Dutch risk adjustment scheme, these distance variables are treated as S-type adjusters in order to compensate for unmeasured health status differences and/or medical supply factors for which the insurers can not be held responsible.

114. Note that these are not geographical regions, but the clusters capture residual cost variation (i.e. variation conditional on the influences of age, sex, eligibility, PCG and DCG) that can be explained by distances to nearest GP and hospital a.o. (see Section 3.3). The regional clusters are not based on regional patterns in hospital output prices.

Table 6.9: Unadjusted and adjusted REF weights from the REF regression with 2001 N-type adjusters included during the estimation phase. The unadjusted REF weights are copied from Table 6.2.

REF adjusters	REF weights [a]	Adjusted REF weights [b]	Change in weights
Intercept	623	653	30
M 15-24 (reference category)	---	---	---
M 25-34	-209	-215	-6
M 35-44	-316	-325	-9
M 45-54	464	462	-2
M 55-64	380	377	-3
M 65-74	1493	1482	-11
M 75-84	2796	2780	-16
M >=85	1598	1585	-13
F 15-24	-109	-120	-11
F 25-34	345	339	-6
F 35-44	0	-6	-6
F 45-54	194	190	-4
F 55-64	274	268	-6
F 65-74	1003	989	-14
F 75-84	2289	2271	-18
F >=85	1662	1634	-28
Disabled	1437	1434	-3
Employed (reference category)	---	---	---
Social welfare	211	188	-23
Unemployed	214	215	1
Retired	341	343	2
Self-Employed	-197	-195	2
Region 1	457	224	-233
Region 2	262	54	-208
Region 3	138	-69	-207
Region 4	239	32	-207
Region 5	37	-156	-193
Region 6	-121	-295	-174
Region 7	-31	-197	-166
Region 8	-164	-265	-101
Region 9	29	-89	-118
Region 10 (reference category)	---	---	---
No PCG (reference category)	---	---	---
Asthma/COPD	1883	1881	-2
Epilepsy	1803	1799	-4

REF adjusters	REF weights [a]	Adjusted REF weights [b]	Change in weights
Crohn/Colitis Ulcerosa	1098	1089	-9
Cardiac disease	2001	2005	4
Rheumatism	3848	3848	0
Parkinson	3199	3177	-22
Diabetes (Type I)	3366	3367	1
Transplantation	7791	7796	5
Cystic fibrosis	3823	3829	6
Neuromuscular disorder	8030	8032	2
HIV/Aids	11895	11877	-18
Renal disease/ESRD	20748	20722	-26
No DCG (reference category)	---	---	---
DCG01	1356	1350	-6
DCG02	6319	6316	-3
DCG03	3565	3573	8
DCG04	5591	5592	1
DCG05	4262	4262	0
DCG06	7820	7855	35
DCG07	6038	6046	8
DCG08	8869	8887	18
DCG09	7983	8014	31
DCG10	18152	18157	5
DCG11	12626	12605	-21
DCG12	9050	9037	-13
DCG13	77982	77953	-29
Hospital output price 2002 (in 100 euro)	---	33	33
Distance to nearest GP 2001 (in km)	---	-112	-112
Distance to nearest hospital 2001 (in km)	---	-5	-5
R^2_{ADJ}	17.93%	17.93%	

Note: PCGs and DCGs are both rank ordered. PCG01=Asthma/COPD, PCG02=Epilepsy, PCG03=Crohn/Colitis Ulcerosa, PCG04=Cardiac disease, PCG05=Rheumatism, PCG06=Parkinson, PCG07=Diabetes (Type I), PCG08=Transplantation, PCG09=Cystic fibrosis, PCG10=Neuromuscular disorder, PCG11=HIV/Aids, PCG12=Renal disease/ESRD.

[a] These estimates result after the estimation of REF equation (2.1).

[b] These estimates result after the estimation of equation (2.1').

for enrollees living in the fifth regional cluster of ZIP codes and especially not for those living in the tenth cluster. Note that, contrary to the assumptions made in this study, in practice the three N-type adjusters are defined as S-type adjusters in the 2004 Dutch REF model. Therefore, to the extent that REF predicted costs

Table 6.10: REF predicted costs compared to normative costs 2002 with and without a correction of the REF weights for omitted variables bias, for subgroups of survey respondents defined by the REF adjusters from the REF equation (2.1): ten regional clusters of 2001 ZIP codes

		REF predicted costs – normative costs		REF predicted costs / normative costs	
Subgroups of enrollees	**Size f subgroup**	**REF weights**	**Adjusted REF weights**	**REF weights**	**Adjusted REF weights**
Region 1	7.2%	245 *	190 *	1.136 *	1.105 *
Region 2	20.6%	84	53	1.044	1.028
Region 3	9.5%	44	16	1.028	1.010
Region 4	9.2%	89 **	61	1.051 **	1.034
Region 5	14.6%	-39	-55	0.978	0.969
Region 6	9.9%	-190 *	-186 *	0.893 *	0.895 *
Region 7	16.7%	-94 *	-80 **	0.946 *	0.954 **
Region 8	2.8%	-171 *	-93	0.886 *	0.938
Region 9	3.3%	-80	-19	0.951	0.988
Region 10	6.1%	3	182 *	1.002	1.108 *
Total	100.0%	0	0	1.000	1.000

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

Table 6.11: REF predicted costs compared to normative costs 2002 with and without an adjustment of the REF weights for omitted variables bias, for subgroups of survey respondents defined by the REF adjusters from the REF equation (2.1): insurance eligibility

		REF predicted – normative costs		REF predicted / normative costs	
Subgroups of enrollees	**Size of subgroup**	**Unadjusted REF weights**	**Adjusted REF weights**	**Unadjusted REF weights**	**Adjusted REF weights**
Disabled	9.0%	420 *	412 *	1.151 *	1.148 *
Employed	59.5%	-24 **	-21	0.976 **	0.978
Social welfare	4.1%	-327 *	-364 *	0.838 *	0.819 *
Unemployed	4.2%	-106	-100	0.938	0.941
Retired	20.5%	-5	-10	0.998	0.997
Self-Employed	2.8%	-162 *	-138 **	0.838 *	0.862 **
Total	100.0%	0	-1	1.000	1.000

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

are determined by these N-type adjusters, their effects are removed partially by application of the omitted variables approach in this study.

Table 6.11 shows that the gap between REF predicted costs and normative costs hardly changes for subgroups defined by the 2004 Dutch REF adjuster eligibility if the N-type adjusters are taken into account during the estimation phase. The

omitted variables approach appears to only slightly remove the cost variation caused by N-type risk factors, with respect to the subgroup of people being on social welfare it even worsens the undercompensation.

According to Table A6.3 in Appendix A6.2, footnote a, a weighted average of (the absolute value of) the subgroup statistic given by equation (2.6) for all subgroups defined by the S-type adjusters is equal to 201. This figure can be compared to the figure of 687 that would result in the absence of any cross-subsidies between the healthy and the sick insured people.[115] The risk-adjusted premium subsidies induced by the 2004 Dutch REF equation can therefore be estimated to be in line with the policy goals of the Dutch government up to an extent of (1-201/687) x 100% = 70.7%. This performance outcome is slightly worse than the 71.2% figure that holds for the REF equation with unadjusted REF weights. Apparently, a removal of the N-type bias from the REF weights following the omitted variables approach at the same time reduces the amount of S-type cost variation that is captured by the REF adjusters.

Table 6.12 is added as an illustration of the importance of quality induced cost differences for which insurers are held responsible and which are not captured by the usual REF approach. The quality measure is derived from a 2004 Dutch newspaper publication concerning one hundred hospitals (AD 2003). At the time of publication, it was for the first time ever that quality measures were made public in such a way that a patient can compare performances of all Dutch hospitals at a glance.

Table 6.12: REF regression weights, given the N-type adjusters hospital output price, distance to nearest GP, distance to nearest hospital and hospital quality are included during the estimation phase (regression weights with respect to other REF adjusters are not tabulated).

REF adjusters	**Adjusted REF weights** [a]
Hospital output price 2002 (in 100 euro)	39
Distance to nearest GP 2001 (in km)	-114
Distance to nearest hospital 2001 (in km)	-5
Hospital quality 2001 (0-100 scale)	-5
R^2_{ADJ}	17.93%

[a] These estimates result after the estimation of equation (2.1'). They are not statistically significantly different from zero. Adjusted REF predicted costs are derived following equation (2.2') given the weights tabulated here.

115. The figure of 687 can be found in Table A6.1, footnote a.

For ranking the hospitals, the Dutch Health Care Inspectorate (IGZ), the Dutch Association of Hospitals (NVZ), and the Dutch Society of Medical Specialists (OMS) constructed 26 quality indicators. Patient experiences with hospital care delivery are not taken into account. The hospital scores are made public on a voluntary basis by the medical clinics of hospitals themselves in order to inform patients about their performance. The Dutch newspaper assigned points to the hospitals based on the reported values of the 26 quality indicators. More points can be gained the more closely an indicator is related to patients' health. The maximum overall amount of quality points to be gained is 54, the actual score ranges from 44 down to 9 points.

A few hospitals only published a limited amount of information, such that ultimately no points are gained for one or more indicators and a low ranking place may be the result. Although this does not necessarily mean that the hospitals deliver bad health care, such a low ranking may be justified under the assumption that hospital management will only be able to judge the effectiveness of their policies and undertake adequate action if such information is available.

In Table 6.12 hospital quality 2001 is defined as a weighted average of individual hospital quality scores per four-digit ZIP code, where the weights are the number of 2001 outpatient hospital contacts of Agis enrollees living in a ZIP code.[116] Zero points are assigned to ten hospitals because of limited information. These ten hospitals are not taken into account in the calculation of the hospital quality score per ZIP code, that is their weight is set to zero. This amounts to 11.5% of all 2001 outpatient contacts that Agis enrollees had.

Table 6.12 shows that for enrollees living in a four-digit ZIP code region in which hospitals operate that have a one point higher quality score than for enrollees living in another region, REF predicted costs are five euro lower, ceteris paribus. The conclusions drawn earlier with respect to region (Table 6.10) and eligibility (Table 6.11) also hold qualitatively if hospital quality 2001 is added as an N-type adjuster (results not presented).

In conclusion, the deviations from normative costs are reduced for most of the subgroups defined by the REF adjusters eligibility and region if the omitted variables approach is applied. However, in general the effects appear to be rather limited, at least given the specific choice of N-type adjusters in this study. Therefore, the undesired compensation for cost variation caused by the N-type risk factors will be only partly removed from the risk-adjusted premium subsidies if

116. These weights are based on outpatient contacts of all Agis enrollees present in the 2001 sickness fund population.

based on REF predicted costs with weights adjusted under the omitted variable approach. Furthermore, the REF equation appears to do a slightly worse job in meeting the policy goals of the Dutch government if the adjusted REF weights are applied instead of the unadjusted REF weights. In Section 6.3, it will be shown that the cost variation caused by N-type risk factors can be removed completely from the risk-adjusted premium subsidies if REF weights are applied that are adjusted following the normative approach developed in this study.

6.3 A NORMATIVE ADJUSTMENT OF THE REF WEIGHTS

In Section 6.1, REF predicted costs were shown not to fully satisfy the criterion of effectiveness. In Section 6.2 the REF weights are adjusted by the omitted variables approach in order to remove cost variation caused by N-type risk factors from the risk-adjusted premium subsidies which are based on REF predicted costs. An alternative solution is to adjust the REF weights by regressing normative costs instead of observed costs on the feasible set of REF adjusters. Adjusted REF weights are thus determined by estimation of equation (2.7). Adjusted REF predicted costs are derived following equation (2.8) by applying the adjusted REF weights. In this way, the subgroups defined by the REF adjusters are cross-subsidized for cost variation caused by S-type risk factors alone.

Table 6.13 shows the amount of change in the REF weights due to the application of the normative adjustment procedure. Remember that, in terms of Figure 2.2, a change in weights represents the difference in slopes of the theoretical and observed relationship between observed costs and the REF adjusters (ceteris paribus). In other words, it is a precise estimate of the bias in the unadjusted REF weights that is caused by N-type risk factors. This bias is completely removed if the adjusted REF weights are applied.[117] For example, the necessary change in REF weights with respect to disabled enrollees equals the difference between the adjusted and unadjusted REF weights, i.e. 950 minus 1437 = -487 euro. This means that to the extent of 487 euro, average REF predicted costs for these enrollees (ceteris paribus) must be attributed to N-type risk factors and can be removed from the unadjusted REF weight completely by applying the normative adjustment procedure. As a comparison, Table 6.9 showed that the REF weights are adjusted by 3 euro only if the omitted variables approach were to be applied

117. Note that estimated standard errors are not presented in this table. For the purpose of risk adjustment, only estimated means are relevant in an economic sense.

Table 6.13: Unadjusted and normatively adjusted REF weights as described by equation (2.8). The unadjusted REF weights are copied from Table 6.2.

REF adjusters	REF weights [a]	Adjusted REF weights [b]	Change in weights
Intercept	623	632	9
M 15-24 (reference category)	---	---	---
M 25-34	-209	-217	-7
M 35-44	-316	-311	6
M 45-54	464	537	73
M 55-64	380	494	114
M 65-74	1493	1400	-93
M 75-84	2796	2684	-113
M >=85	1598	1483	-115
F 15-24	-109	-112	-4
F 25-34	345	351	6
F 35-44	0	11	11
F 45-54	194	214	20
F 55-64	274	301	27
F 65-74	1003	917	-86
F 75-84	2289	2193	-96
F >=85	1662	1561	-101
Disabled	1437	950	-487
Employed (reference category)	---	---	---
Social welfare	211	522	311
Unemployed	214	262	48
Retired	341	384	42
Self-Employed	-197	-80	117
Region 1	457	202	-255
Region 2	262	169	-93
Region 3	138	85	-53
Region 4	239	152	-87
Region 5	37	73	36
Region 6	-121	66	187
Region 7	-31	57	89
Region 8	-164	15	180
Region 9	29	99	70
Region 10 (reference category)	---	---	---
No PCG (reference category)	---	---	---
PCG01	1883	2108	225
PCG02	1803	1767	-36
PCG03	1098	412	-686
PCG04	2001	2227	226

REF adjusters	REF weights [a]	Adjusted REF weights [b]	Change in weights
PCG05	3848	3873	25
PCG06	3199	3110	-89
PCG07	3366	3500	134
PCG08	7791	7801	10
PCG09	3823	3899	75
PCG10	8030	8411	381
PCG11	11895	11946	51
PCG12	20748	20823	75
No DCG (reference category)	---	---	---
DCG01	1356	1475	119
DCG02	6319	5211	-1108
DCG03	3565	4442	877
DCG04	5591	6138	548
DCG05	4262	3605	-657
DCG06	7820	7764	-56
DCG07	6038	5550	-487
DCG08	8869	8532	-337
DCG09	7983	7894	-90
DCG10	18152	18125	-27
DCG11	12626	12584	-42
DCG12	9050	10947	1897
DCG13	77982	78036	54
R^2_{ADJ}	17.93%	90.73%	

Note: Both PCGs and DCGs are rank-ordered. PCG01=Asthma/COPD, PCG02=Epilepsy, PCG03=Crohn/Colitis Ulcerosa, PCG04=Cardiac disease, PCG05=Rheumatism, PCG06=Parkinson, PCG07=Diabetes (Type I), PCG08=Transplantation, PCG09=Cystic fibrosis, PCG10=Neuromuscular disorder, PCG11=HIV/Aids, PCG12=Renal disease/ESRD.
[a] These estimates result after the estimation of REF equation (2.1). REF predicted costs are derived following equation (2.2) given the unadjusted REF weights tabulated here.
[b] These estimates result after the estimation of equation (2.7). Adjusted REF predicted costs are derived following equation (2.8) given the adjusted REF weights tabulated here.

instead. Furthermore, the omitted variables approach would remove only 2 euro out of the 117 euro of the change in REF weights which appears to be necessary to avoid undercompensation of self-employed enrollees.[118]

Table 6.9 showed that, in order to avoid compensation for regional cost variation caused by N-type risk factors, under the omitted variables approach all REF

118. The estimated percentage of explained variance (R^2_{ADJ}) in equation (2.8) is much larger than in case of the normative equation (2.4), because at the individual level the variance of normative costs (i.e. the dependent variable in equation (2.7)) is much smaller than that of observed costs in equation (i.e. the dependent variable in equation (2.3)).

weights should be adjusted downwards relative to the REF weight for the subgroup of insured living in the tenth region. However, Table 6.13 shows that only the REF weights with respect to the first four regional clusters should be adjusted downwards under the normative approach. The REF weights for regions 5-9 have to be adjusted upwards relative to the tenth region instead of downwards as opposed to the omitted variables approach. Note that the adjustment downwards is largest for the first regional cluster under both approaches. The REF weights associated with 9 out of 13 PCGs are adjusted upwards and for 8 out of 13 DCGs they are adjusted downwards under the normative approach.

In Table 6.14, both REF predicted costs with unadjusted REF weights and REF predicted costs with adjusted REF weights are compared to normative costs. REF predicted costs with unadjusted REF weights are already shown in Table 6.3, from which it appeared that there are no differences with normative costs as the REF adjusters age and gender are also included in the normative equation. As age and gender are also included in regression (2.7) to find the adjusted REF weights

Table 6.14: REF predicted costs compared to normative costs 2002 before and after a normative adjustment of the REF weights, for subgroups of survey respondents defined by the REF adjusters from the REF equation (2.1): age and gender [a]

		REF predicted – normative costs		REF predicted / normative costs	
Subgroups of enrollees	**Size of subgroup**	**REF weights**	**Adjusted REF weights**	**REF weights**	**Adjusted REF weights**
M 15-24	4.2%	0	0	1.000	1.000
M 25-34	7.1%	0	0	1.000	1.000
M 35-44	6.9%	0	0	1.000	1.000
M 45-54	6.3%	0	0	1.000	1.000
M 55-64	5.7%	0	0	1.000	1.000
M 65-74	5.0%	0	0	1.000	1.000
M 75-84	2.3%	0	0	1.000	1.000
M >=85	0.3%	0	0	1.000	1.000
F 15-24	6.2%	0	0	1.000	1.000
F 25-34	11.3%	0	0	1.000	1.000
F 35-44	12.8%	0	0	1.000	1.000
F 45-54	11.3%	0	0	1.000	1.000
F 55-64	8.9%	0	0	1.000	1.000
F 65-74	7.2%	0	0	1.000	1.000
F 75-84	4.0%	0	0	1.000	1.000
F >=85	0.7%	0	0	1.000	1.000
Total	100.0%	0	0	1.000	1.000

[a] The adjusted REF weights are described mathematically in equation (2.8).

Table 6.15: REF predicted costs compared to normative costs 2002 before and after a normative adjustment of the REF weights, for subgroups of survey respondents defined by the REF adjusters from the REF equation (2.1): insurance eligibility

		REF predicted – normative costs		REF predicted / normative costs	
Subgroups of enrollees	**Size of subgroup**	**REF weights**	**Adjusted REF weights**	**REF weights**	**Adjusted REF weights**
Disabled	9.0%	420 *	0	1.151 *	1.000
Employed	59.5%	-24 **	0	0.976 **	1.000
Social welfare	4.1%	-327 *	0	0.838 *	1.000
Unemployed	4.2%	-106	0	0.938	1.000
Retired	20.5%	-5	0	0.998	1.000
Self-Employed	2.8%	-162 *	0	0.838 *	1.000
Total	100.0%	0	0	1.000	1.000

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

according to the normative adjustment procedure, equality between REF predicted costs and normative costs also holds if a normative adjustment of the REF weights is applied. Therefore, equality between normative costs and adjusted REF predicted costs for the subgroups defined by the REF adjusters age and gender holds by construction.

Table 6.15 shows the difference between unadjusted and adjusted REF predicted costs for subgroups defined by the REF adjuster eligibility. The pattern observed with respect to unadjusted REF predicted costs is already discussed in the context of Table 6.7. Furthermore, adjusted REF predicted costs coincide with normative costs as the subgroups defined by the REF adjuster eligibility are used as explanatory variable for normative costs in equation (2.8) in order to derive adjusted REF predicted costs. As a consequence, if these adjusted REF weights are applied to eligibility instead of the unadjusted REF weights, eligibility induces S-type cost variation alone. Although the results are not tabulated here, the same conclusion can be drawn with respect to the other REF adjusters region, the PCGs and the DCGs.

According to Table A6.3 in Appendix A6.3, footnote a, a weighted average of (the absolute value of) the subgroup statistic given by equation (2.6) for all subgroups defined by the S-type adjusters is equal to 209. This figure can be compared to the figure of 687 that would result in the absence of any cross-subsidies between the healthy and the sick insured people.[119] The risk-adjusted premium subsidies

119. The figure of 687 can be found in Table A6.1, footnote a.

induced by the 2004 Dutch REF equation can therefore be estimated to be in line with the policy goals of the Dutch government up to an extent of (1-209/687) x 100% = 69.6%. This performance outcome is slightly worse than the 71.2% figure that holds for the REF equation with unadjusted REF weights, even more so than when adjusting the REF weights by the omitted variables approach. Apparently, a removal of the N-type bias from the REF weights following the normative adjustment procedure at the same time reduces the amount of S-type cost variation that is captured by the REF adjusters.

To summarize this section, it must be concluded that a normative adjustment of the REF weights generates REF predicted costs that are adjusted such that these are exactly in line with normative costs if the risk-adjusted premium subsidies are based on subgroups defined by the REF adjusters. Therefore, risk-adjusted premium subsidies based on normatively adjusted REF predicted costs are preferred to risk-adjusted premium subsidies based on unadjusted REF predicted costs in practice. However, the REF equation appears to do a slightly worse job in meeting the policy goals of the Dutch government if the adjusted REF weights are applied instead of the unadjusted REF weights, even worse than when applying the omitted variables approach to adjust the REF weights.

6.4 Conclusions

REF predicted costs do not coincide with normative costs for the subgroups defined by the S-type adjusters. This is a result by construction. Because of feasibility restrictions, REF predicted costs are based on observed costs for the subgroups defined by a limited set of REF adjusters instead of the broad set of S-type adjusters. Therefore, given the definition of normative costs in this study, it must be concluded that the 2004 Dutch REF equation does not (fully) satisfy the criterion of effectiveness of risk-adjusted premium subsidies. At one extreme, REF predicted costs fall short of normative costs by 27.4% for those with three or more self-reported OECD limitations or chronic conditions. At the other extreme, REF predicted costs are 54.1% above normative costs for those insured people with top quartile physical functioning (PF) scale scores. In conclusion, the risk-adjusted premium subsidies induced by the 2004 Dutch REF equation are estimated to be in line with the policy goals of the Dutch government up to an extent of (1-198/687) x 100% = 71.2%.

In practice, risk-adjusted premium subsidies are based on the subgroups defined by the REF adjusters. Deviations from normative costs for such a subgroup

may be interpreted as bias caused by N-type risk factors. This is due to the fact that the normative costs are derived from a reduced form specification of the normative equation (2.3): N-type cost variation that is caused by correlated S-type and N-type risk factors is also included in the definition of normative costs. Based on the normative approach developed in this study, REF predicted costs for the subgroup of disabled enrollees appear to be 15.1% above normative costs on average, whereas for enrollees on social welfare and for self-employed enrollees REF predicted costs are 16.2% below normative costs on average. Furthermore, REF predicted costs for enrollees living in the first out of ten regional clusters of ZIP codes are 13.6% above normative costs. The 6th, 7th and 8th regional clusters contain enrollees for whom REF predicted costs lie between 5.4% and 11.4% below normative costs on average. From these results it is concluded that if the risk-adjusted premium subsidies to these subgroups are based on REF predicted costs induced by the 2004 Dutch REF equation, then these subsidies will conflict with the policy goals of the Dutch government to a certain extent.

Given the feasible set of REF adjusters, an omitted variables approach may be applied in order to remove this undesirable bias from the REF weights associated with the REF adjusters eligibility and region. However, the adjustment that follows by application of this omitted variables approach appears to be rather limited, at least given the specific choice of the N-type adjusters applied in this study. As an alternative approach to remove the undesirable bias, a normative adjustment of the REF weights is proposed. Basically, this approach means that normative costs instead of observed costs are regressed on the REF adjusters for the survey respondents. By construction, for each subgroup defined by the REF adjusters, normatively adjusted REF predicted costs are identical to normative costs on average. As the bias is therefore completely removed from the REF weights, it is recommended to use normatively adjusted REF weights instead of unadjusted REF weights for the calculation of risk-adjusted premium subsidies in practice. However, there is also a tradeoff to be made: irrespective of the procedure that will be chosen to adjust the REF weights, a removal of the N-type bias from the REF weights appears to slightly reduce the amount of S-type cost variation that is captured by the REF adjusters, at least in this study sample. Note that it is possible to apply the normatively adjusted REF weights to all 16 million Dutch insured, not only to the 18,617 survey respondents who are available for the estimation of normative costs in this study.

In Chapter Seven, alternative specifications of the 2004 Dutch REF model are tested for the extent to which these specifications satisfy the criterion of effectiveness. The purpose of these exercises is to compare REF predicted costs based on

these alternative risk equalization models with REF predicted costs from the 2004 Dutch REF equation, for the subgroups defined by the S-type adjusters. Furthermore, although normatively adjusted weights will not be calculated in Chapter 7 for these alternative risk equalization models, it is recommended to always use adjusted weights instead of unadjusted weights in practice.

APPENDIX A6.1: A NORMATIVE TEST OF THE PRE-2004 DUTCH REF EQUATIONS

In Section 6.1 a normative test of the 2004 Dutch REF equation is presented. The REF adjusters included in REF equation (2.1) are not all present since the implementation of the first Dutch risk equalization model in 1991. REF adjusters are added successively over the years as they became available for all relevant insurance members. For example, only since 2004 the DCGs are included.[120] In this section, REF predicted costs from four specifications are compared to normative costs for subgroups of survey respondents defined by the S-type adjusters. The results for the 2004 Dutch REF equation are presented in the last column, which are also shown in Tables 6.3-6.6 of Section 6.1.

Table A6.1: REF predicted costs compared to normative costs 2002 given four variants of the REF equation, for subgroups of survey respondents defined by S-type adjusters from the normative equation (2.3): SF-36 scores, number of self-reported OECD limitations and number of self-reported chronic conditions [a]

		REF predicted – normative costs			
Subgroups	**Size of subgroup**	**No adjusters**	**Demo**	**Demo + PCGs**	**Demo + PCGs + DCGs**
Q1 PF scores	25.0%	-1989 *	-1113 *	-877 *	-752 *
Q1 RP scores	25.0%	-1746 *	-1041 *	-798 *	-621 *
Q1 BP scores	25.0%	-1385 *	-857 *	-726 *	-628 *
Q1 GH scores	25.0%	-1758 *	-1102 *	-804 *	-599 *
Q1 VT scores	25.0%	-1102 *	-755 *	-579 *	-419 *
Q1 SF scores	25.0%	-1200 *	-801 *	-657 *	-514 *
Q1 RE scores	25.0%	-934 *	-517 *	-407 *	-318 *
Q1 MH scores	25.0%	-577 *	-356 *	-289 *	-221 *
Q2 PF scores	25.0%	51	219 *	172 *	124 *
Q2 RP scores	25.0%	217 *	156 *	101 *	77 *
Q2 BP scores	25.0%	73 **	60	40	40
Q2 GH scores	25.0%	92 *	167 *	107 *	46 **
Q2 VT scores	25.0%	6	-3	38	-3
Q2 SF scores	25.0%	-81 *	-31	6	1
Q2 RE scores	25.0%	282 *	188 *	144 *	103 *
Q2 MH scores	25.0%	-72	-40	-20	4
Q3 PF scores	25.0%	805 *	421 *	322 *	292 *

120. In Appendix A1.2 of Chapter 1 the sickness funds REF models with respect to the years 1991-2005 are described in more detail.

		REF predicted – normative costs			
Subgroups	**Size of subgroup**	**No adjusters**	**Demo**	**Demo + PCGs**	**Demo + PCGs + DCGs**
Q3 RP scores	25.0%	697 *	478 *	374 *	288 *
Q3 BP scores	25.0%	612 *	341 *	278 *	252 *
Q3 GH scores	25.0%	630 *	430 *	320 *	225 *
Q3 VT scores	25.0%	458 *	278 *	194 *	163 *
Q3 SF scores	25.0%	565 *	401 *	320 *	250 *
Q3 RE scores	25.0%	165 *	154 *	128 *	118 *
Q3 MH scores	25.0%	313 *	170 *	131 *	93 *
Q4 PF scores	25.0%	1133 *	473 *	383 *	336 *
Q4 RP scores	25.0%	832 *	406 *	322 *	256 *
Q4 BP scores	25.0%	699 *	456 *	408 *	336 *
Q4 GH scores	25.0%	1035 *	506 *	377 *	328 *
Q4 VT scores	25.0%	637 *	480 *	347 *	259 *
Q4 SF scores	25.0%	716 *	431 *	331 *	263 *
Q4 RE scores	25.0%	487 *	175 *	135 *	97 *
Q4 MH scores	25.0%	336 *	227 *	178 *	123 *
Number of self-reported OECD limitations					
0	77.8%	536 *	289 *	244 *	204 *
1	10.3%	-1157 *	-455 *	-396 *	-370 *
2	4.9%	-2201 *	-1185 *	-1003 *	-916 *
3+	5.2%	-3524 *	-2250 *	-1870 *	-1445 *
Imputed	1.8%	-284	-129	-108	-13
Number of self-reported chronic conditions					
0	90.4%	339 *	217 *	147 *	93 *
1	8.1%	-2839 *	-1769 *	-1228 *	-778 *
2	1.3%	-4782 *	-3225 *	-1970 *	-1268 *
3+	0.2%	-7428 *	-5777 *	-4055 *	-2511 *
Total	100.0%	0	0	0	0

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

[a] The weighted averages of the absolute values of the differences between REF predicted costs and normative costs for the tabulated subgroups in the columns "No adjusters", "Demo", "Demo + PCGs" and "Demo + PCGs + DCGs" are equal to 680, 410, 320 and 251, respectively. If the subgroups defined by the S-type adjusters age, gender, PCGs and DCGs are also taken into account (see Table A6.2) then these figures are equal to 687, 392, 279 and 198.

Table A6.2: REF predicted costs compared to normative costs 2002 given four variants of the REF equation, for subgroups of survey respondents defined by S-type adjusters from the normative equation (2.3): age and gender, PCGs and DCGs

		REF predicted – normative costs			
Subgroups	**Size of subgroup**	**No adjusters**	**Demo**	**Demo + PCGs**	**Demo + PCGs + DCGs**
M 15-24	4.2%	935 *	0	0	0
M 25-34	7.1%	1062 *	0	0	0
M 35-44	6.9%	992 *	0	0	0
M 45-54	6.3%	-148	0	0	0
M 55-64	5.7%	-624 *	0	0	0
M 65-74	5.0%	-2031 *	0	0	0
M 75-84	2.3%	-3291 *	0	0	0
M >=85	0.3%	-1988 *	0	0	0
F 15-24	6.2%	941 *	0	0	0
F 25-34	11.3%	497 *	0	0	0
F 35-44	12.8%	743 *	0	0	0
F 45-54	11.3%	303 *	0	0	0
F 55-64	8.9%	-114 **	0	0	0
F 65-74	7.2%	-1061 *	0	0	0
F 75-84	4.0%	-2509 *	0	0	0
F >=85	0.7%	-1875 *	0	0	0
No PCG	91.2%	372 *	267 *	9	9
PCG01	4.0%	-3322 *	-2422 *	-555 *	-422
PCG02	0.5%	-2725 *	-2168 *	49	-65
PCG03	0.2%	-1957 *	-1820 *	607	482
PCG04	3.1%	-4833 *	-3023 *	-305 **	-261
PCG05	0.3%	-4215 *	-3912 *	-34	-12
PCG06	0.1%	-4697 *	-2859 *	23	-40
PCG07	1.2%	-4463 *	-3528 *	-58	-48
PCG08	0.1%	-10192 *	-9406 *	-195	-208
PCG09	0.0%	-5732 *	-5088 **	0	0
PCG10	0.1%	-9994 *	-8691 *	0	0
PCG11	0.1%	-11529 *	-11814 *	0	0
PCG12	0.0%	-38388 *	-37512 *	0	0
No DCG	97.2%	231 *	201 *	170 *	0
DCG01	0.5%	-3205 *	-1938 *	-1561 *	-185
DCG02	0.6%	-7415 *	-6436 *	-5471 *	-147
DCG03	0.5%	-6927 *	-6091 *	-5319 *	-1482
DCG04	0.5%	-9320 *	-8247 *	-7670 *	-1109
DCG05	0.3%	-6540 *	-5566 *	-4231 *	261
DCG06	0.1%	-9408 *	-8117 *	-7506 *	-98

		REF predicted – normative costs			
Subgroups	**Size of subgroup**	**No adjusters**	**Demo**	**Demo + PCGs**	**Demo + PCGs + DCGs**
DCG07	0.3%	-8786 *	-8103 *	-5344 *	78
DCG08	0.2%	-11683 *	-10022 *	-8310 *	123
DCG09	0.0%	-9304 *	-8136 *	-7589 *	0
DCG10	0.1%	-21883 *	-20457 *	-16398 *	-290
DCG11	0.1%	-13341 *	-12865 *	-12236 *	0
DCG12	0.1%	-22139 *	-21218 *	-17937 *	-2717
DCG13	0.0%	-82371 *	-81030 *	-74577 *	0
Total	100.0%	0	0	0	0

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

Appendix A6.2: A supplement to Section 6.2 and Section 6.3

In Tables 6.3 and 6.4 the REF predicted costs are compared to normative costs before and after adjustment of the REF weights. The REF weights are adjusted following the omitted variables approach and the normative adjustment procedure. The column "REF weights" is equal to the column "Demo + PCGs + DCGs" in Tables A6.1 and A6.2.

Table A6.3: REF predicted costs compared to normative costs 2002 before and after adjustment of the REF weights, for subgroups of survey respondents defined by S-type adjusters from the normative equation (2.3): SF-36 scores, number of self-reported OECD limitations and number of self-reported chronic conditions

		REF predicted – normative costs			
Subgroups	**Size of subgroup**	**No adjusters**	**REF weights**	**Adjusted REF weights – omitted variables approach**	**Adjusted REF weights – normative adj. approach**
Q1 PF scores	25.0%	-1989 *	-752 *	-761 *	-787 *
Q1 RP scores	25.0%	-1746 *	-621 *	-627 *	-666 *
Q1 BP scores	25.0%	-1385 *	-628 *	-637 *	-669 *
Q1 GH scores	25.0%	-1758 *	-599 *	-609 *	-642 *
Q1 VT scores	25.0%	-1102 *	-419 *	-428 *	-455 *
Q1 SF scores	25.0%	-1200 *	-514 *	-523 *	-554 *
Q1 RE scores	25.0%	-934 *	-318 *	-327 *	-342 *
Q1 MH scores	25.0%	-577 *	-221 *	-230 *	-246 *
Q2 PF scores	25.0%	51	124 *	123 *	115 *
Q2 RP scores	25.0%	217 *	77 *	74 *	81 *

		REF predicted – normative costs			
Subgroups	**Size of subgroup**	**No adjusters**	**REF weights**	**Adjusted REF weights – omitted variables approach**	**Adjusted REF weights – normative adj. approach**
Q2 BP scores	25.0%	73 **	40	39	41
Q2 GH scores	25.0%	92 *	46 **	45 **	46 **
Q2 VT scores	25.0%	6	-3	-5	-1
Q2 SF scores	25.0%	-81 *	1	0	4
Q2 RE scores	25.0%	282 *	103 *	103 *	106 *
Q2 MH scores	25.0%	-72	4	3	7
Q3 PF scores	25.0%	805 *	292 *	296 *	312 *
Q3 RP scores	25.0%	697 *	288 *	292 *	312 *
Q3 BP scores	25.0%	612 *	252 *	253 *	265 *
Q3 GH scores	25.0%	630 *	225 *	227 *	242 *
Q3 VT scores	25.0%	458 *	163 *	166 *	179 *
Q3 SF scores	25.0%	565 *	250 *	253 *	268 *
Q3 RE scores	25.0%	165 *	118 *	124 *	134 *
Q3 MH scores	25.0%	313 *	93 *	96 *	106 *
Q4 PF scores	25.0%	1133 *	336 *	339 *	361 *
Q4 RP scores	25.0%	832 *	256 *	258 *	273 *
Q4 BP scores	25.0%	699 *	336 *	342 *	362 *
Q4 GH scores	25.0%	1035 *	328 *	333 *	354 *
Q4 VT scores	25.0%	637 *	259 *	263 *	276 *
Q4 SF scores	25.0%	716 *	263 *	266 *	282 *
Q4 RE scores	25.0%	487 *	97 *	97 *	102 *
Q4 MH scores	25.0%	336 *	123 *	128 *	134 *
Number of self-reported OECD limitations					
0	77.8%	536 *	204 *	206 *	215 *
1	10.3%	-1157 *	-370 *	-376 *	-397 *
2	4.9%	-2201 *	-916 *	-926 *	-963 *
3+	5.2%	-3524 *	-1445 *	-1459 *	-1503 *
Imputed	1.8%	-284	-13	-20	-1
Number of self-reported chronic conditions					
0	90.4%	339 *	93 *	93 *	97 *
1	8.1%	-2839 *	-778 *	-782 *	-802 *
2	1.3%	-4782 *	-1268 *	-1273 *	-1303 *
3+	0.2%	-7428 *	-2511 *	-2525 *	-2700 *
Total	100.0%	0	0	0	0

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

[a] The weighted averages of the absolute values of the differences between REF predicted costs and normative costs for the tabulated subgroups in the columns "No adjusters", "REF weights", "Adjusted REF weights – omitted variables approach" and "Adjusted REF weights – normative adj. approach" are equal to 680, 251, 254 and 267, respectively. If the subgroups defined by the S-type adjusters age, gender, PCGs and DCGs are also taken into account (see Table A6.4) then these figures are equal to 687, 198, 201 and 209.

Table A6.4: REF predicted costs compared to normative costs 2002 before and after adjustment of the REF weights, for subgroups of survey respondents defined by S-type adjusters from the normative equation (2.3): age and gender, PCGs and DCGs

		REF predicted – normative costs			
Subgroups	**Size of subgroup**	**No adjusters**	**REF weights**	**Adjusted REF weights – omitted variables approach**	**Adjusted REF weights – normative adj. approach**
M 15-24	4.2%	935 *	0	13	0
M 25-34	7.1%	1062 *	0	-1	0
M 35-44	6.9%	992 *	0	-16	0
M 45-54	6.3%	-148	0	7	0
M 55-64	5.7%	-624 *	0	5	0
M 65-74	5.0%	-2031 *	0	3	0
M 75-84	2.3%	-3291 *	0	-2	0
M >=85	0.3%	-1988 *	0	-4	0
F 15-24	6.2%	941 *	0	-7	0
F 25-34	11.3%	497 *	0	1	0
F 35-44	12.8%	743 *	0	-1	0
F 45-54	11.3%	303 *	0	3	0
F 55-64	8.9%	-114 **	0	4	0
F 65-74	7.2%	-1061 *	0	-5	0
F 75-84	4.0%	-2509 *	0	-19	0
F >=85	0.7%	-1875 *	0	-15	0
No PCG	91.2%	372 *	9	8	0
PCG01	4.0%	-3322 *	-422	-432 *	-271 *
PCG02	0.5%	-2725 *	-65	-67	-159
PCG03	0.2%	-1957 *	482	474	-122
PCG04	3.1%	-4833 *	-261	-259	-139
PCG05	0.3%	-4215 *	-12	7	-42
PCG06	0.1%	-4697 *	-40	-52	-181
PCG07	1.2%	-4463 *	-48	-52	-16
PCG08	0.1%	-10192 *	-208	-215	-35
PCG09	0.0%	-5732 *	0	-2	0
PCG10	0.1%	-9994 *	0	-6	0
PCG11	0.1%	-11529 *	0	-36	0

		REF predicted – normative costs			
Subgroups	**Size of subgroup**	**No adjusters**	**REF weights**	**Adjusted REF weights – omitted variables approach**	**Adjusted REF weights – normative adj. approach**
PCG12	0.0%	-38388 *	0	-31	0
No DCG	97.2%	231 *	0	-1	0
DCG01	0.5%	-3205 *	-185	-193	-91
DCG02	0.6%	-7415 *	-147	-144	-899 *
DCG03	0.5%	-6927 *	-1482	-1483 *	-737
DCG04	0.5%	-9320 *	-1109	-1101	-709
DCG05	0.3%	-6540 *	261	251	-279
DCG06	0.1%	-9408 *	-98	-59	-218
DCG07	0.3%	-8786 *	78	82	-342
DCG08	0.2%	-11683 *	123	135	-105
DCG09	0.0%	-9304 *	0	19	0
DCG10	0.1%	-21883 *	-290	-291	-236
DCG11	0.1%	-13341 *	0	-23	0
DCG12	0.1%	-22139 *	-2717	-2761	-962
DCG13	0.0%	-82371 *	0	-347	0
Total	100.0%	0	0	0	0

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

Chapter 7

TESTING ALTERNATIVE REF MODEL SPECIFICATIONS FOR EFFECTIVENESS

In Chapter Six, the normative test procedure has been applied to the 2004 Dutch REF equation, given the normative costs derived in Chapter Five. The REF equation appears not to fully satisfy the criterion of effectiveness. The procedure followed there may also prove valuable in order to test alternative specifications of the REF model. In this chapter, illustrations are given of how to apply this procedure on the basis of the research sample in this study. If some specification turns out to better satisfy the criterion of effectiveness than the 2004 Dutch REF model specification, then this must be considered to implement in practice as it improves affordability for the total population of Dutch insured people.

In section 7.1 new risk adjusters are added to the 2004 Dutch REF equation: paramedic diagnostic referral codes of chronic diseases are used as indicators of physical limitations, five types of medical devices as indicators of functional problems, and four types of pharmaceutical drugs acting on the nervous system as indicators of mental diseases. These variables are derived from the 2001 claims data in the Agis sickness fund administration.

In section 7.2 an analogue of the 2004 Dutch risk sharing scheme is applied, which essentially boils down to a 90% retrospective reimbursement of actual health care costs above a threshold of € 12,500.

In Section 7.3 REF weights are estimated in the context of a GLM framework under the assumption of a Gamma error distribution and a log link between REF predicted costs and the REF adjusters.

In Section 7.4 the conclusions are drawn.

7.1 Adding new risk adjusters to the REF equation

The following additional administrative variables are tested for their potential as a new REF adjuster, in addition to those already included in the 2004 Dutch REF equation:

1. Paramedic diagnostic referral codes with a chronic indication;
2. Medical devices;
3. Pharmaceutical drugs acting on the nervous system that are used for more than half a year.

These variables are derived from the claims data in the 2001 Agis sickness fund administration. It should be noted that the classifications presented here are not reviewed by medical experts. Such a medical review is recommended before using these results for implementation.

Paramedic Cost Groups

From the claims data the diagnostic referral codes are retrieved for those Agis members with a chronic indication for paramedic treatment in 2001. These diagnostic referral codes are rank ordered and clustered in order to mitigate the possibilities of discretionary coding behaviour by health care providers.[121]

First, expected costs are calculated for all enrollees in the dataset by applying the regression weights that result from a linear regression of 2002 costs on age and gender interaction terms, and (rank ordered) PCG dummy variables, based on the Agis population of enrollees without any diagnostic referral code.[122] Second, for each subgroup of enrollees with a diagnostic referral code, the difference between average actual costs and these average expected costs is determined per diagnostic referral code. Third, the diagnostic referral code associated with the highest difference between observed and predicted costs is identified and enrollees who belong to this subgroup are removed from the data set. The second and third step are repeated until the data set consists of persons without diagnostic referral codes only. Therefore, an enrollee is associated with the highest ranked diagnostic referral code only. The order of removing the diagnostic referral codes defines the ranking of diagnostic referral codes according to decreasing expected costs.[123]

Note that paramedic chronic conditions are not rank-ordered merely by average observed costs, but by the deviation of observed costs from costs expected for those without any such condition. The purpose of this procedure is to prevent a change in the ranking of the diagnostic referral codes according to their estimated weights, if added as new REF adjusters to the ones already included in the REF equation. This might otherwise occur because of multicollinearity between the paramedic chronic conditions and the REF adjusters already included in the 2004 Dutch REF equation.

Table 7.1 shows the rank ordered paramedic chronic codes. From the last row it appears that 34.53 out of 1000 Agis enrollees have a paramedic diagnostic referral code, with average actual costs € 2121 above the level of expenses that might be expected given the age, sex, and PCG composition of this subgroup of enrollees. It

---

121. Most of the enrollees from Amsterdam received paramedic treatment under the so-called Amsterdam Paramedic Services Model in 2001. This alternative delivery of care process also implied a diagnostic coding system that differed from the national system to some extent, therefore the population used for the rank ordering and clustering does not include these enrollees.

122. The information on the DCGs is only available for the research sample of respondents to the Agis Health Survey 2001. As the rank-ordering procedure is applied to the total Agis sickness fund population, the DCG dummy variables cannot be used for this purpose.

123. See also Lamers and Van Vliet (2003) for a description of this iterative procedure in the context of the construction of the pharmacy-based cost groups (PCGs).

Table 7.1: Rank ordering and clustering of 100 categories of 2001 paramedic diagnostic referral codes (00-99), enrollees belong to one category only (2002 costs, N = 1.0 million enrollees)

Diagnostic referral code	Description	Average actual – expected costs (pipy) [a]	Weighted prevalence per 1000 enrollees [b]	Para-medical cost groups [c]
51	Congenital defects tractus respiratorus	10094 *	0.05	5
76	Spinal cord lesion	9143 *	0.14	5
69	Malignancies without surgery	8559 *	0.11	5
54	COPD	8226 *	0.56	5
14	Inflammations	6772 *	0.10	4
56	Besnier Boeck disease, diffuse interstitial lung disorder, sarcoidose	6528 *	0.02	4
73	Multiple sclerosis/A.L.S./spinales	5644 *	0.60	4
68	Surgery, not locomotor apparatus, not cardio surgery	4805 *	0.25	4
78	Other neurologic conditions	4699 *	0.69	4
48	General vascular dysfunction	4377 *	0.37	4
43	PTCA	4795	0.00	4
77	Neurotrauma	3999 *	0.19	3
00	Amputation	3818 *	0.37	3
41	Hart infarct, myocard-infarct (AMI)	3722 **	0.04	3
71	Cerebellar disorders/encephal	3515 *	0.60	3
90	(Chronic) Rheumatoid arthritis	3448 *	1.81	3
92	Aseptic (poly)arthritis	3364 *	0.94	3
09	Other surgical diseases	3357 *	0.13	3
65	Other, hereditery diseases	3176 *	0.04	3
91	Juvenile rheumatism	3151 *	0.05	3
96	Scleroderma	3060 *	0.02	3
42	Coronary artery bypass operation (CABG)	3046 *	0.03	3
72	Cerebrovascular accident/central pareses	3029 *	1.92	3
99	Other skin diseases	2993 *	0.06	3
24	Osteoporosis	2720 *	0.40	3
46	Lymphatic vessel disease/oedema	2605 *	0.77	3
74	Parkinson/extrapyramidal disorder	2484 *	0.72	3
16	Extreme posture defect	2298 *	0.06	3
44	Heart(valve)surgery	2106	0.01	2
12	Congenital skeleton defect	2100 *	0.59	2
95	Cicatricial tissue	2068 *	0.03	2
28	Sudeck's a(dys)trophy	2036 *	0.66	2
93	Spondylitis ankylopoetica, ankylosing	1833 *	0.42	2
08	Postoperative contracture, atrophie	1777 *	0.30	2

Diagnostic referral code	Description	Average actual – expected costs (pipy) [a]	Weighted prevalence per 1000 enrollees [b]	Para-medical cost groups [c]
01	Articular, except spinal column	1716 *	4.16	2
05	Spinal column	1678 *	1.35	2
94	Other rheumatic- and collagen	1634 *	0.80	2
39	Burns (status after)	1618	0.03	2
02	Bones, except spinal column	1601 *	1.33	2
70	Peripheral nerve disorders	1508 *	1.49	2
79	Psychomotoric retardation	1337 *	2.20	2
32	Luxation, post-traumatic	1283 *	0.30	2
13	Ossification disorder	1198 *	0.06	2
11	Spinal column defect, pelvic	998 *	0.98	1
33	Muscle rupture, tendon rupture, haematoma	924	0.21	1
04	Tendon, muscle, ligament	834 *	1.18	1
36	Fractures	783 *	1.81	1
89	Gynaecology	773	0.61	1
21	Bursitis (not traumatic), capsulitis	738 *	1.89	1
38	Whiplash injury (neck trauma)	683 *	0.91	1
75	HNP with motor deficits	669 *	0.91	1
03	Meniscitis, synovectomy	587 *	1.17	1
10	Aseptic bone necrosis	-27	0.06	---
45	Surgical correction congenital defects	-724	0.00	---
All	---	2121 *	34.53	---

[a] For each subgroup of enrollees associated with a specific diagnostic referral code, the average difference is presented between their actual costs and costs that might be expected given their 2001 age, sex, and (rank ordered) PCG categorization. Expected costs are derived given the estimated coefficients from a linear regression of 2002 costs on 2001 age/sex and (rank ordered) PCG dummy variables, restricted to the subpopulation of enrollees for whom no diagnostic referral code is present in 2001. These calculations are performed on the claims data of the Agis population (exclusive of the former ZAO members), i.e. not on those of the respondents to the Agis Health Survey 2001 alone. Regression weights are applied to correct for the number of months of membership in 2002, partial months being rounded upward.

[b] A member for whom more than one diagnostic referral code was found, is assigned to the subgroup for which the average difference between actual costs and expected costs is largest. Subgroups associated with diagnostic referral codes that are not tabulated have zero prevalence after applying this procedure. Reported prevalences are weighted in order to correct for the number of months of membership in 2002, partial months being rounded upward.

[c] Ward's minimum variance method is applied in order to derive the clusters, taking account of average differences between actual and expected costs, number of observations, and standard deviations. The decision to apply five clusters is based on the outcomes of the pseudo t2 statistic (not reported here).

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

follows that 96.55% of the Agis enrollees form the subgroup for which no diagnostic referral code with a chronic indication is found in the 2001 Agis claims data.

For reasons of model stability[124] and administrative feasibility, the diagnostic referral codes are merged such that five clusters of so-called paramedic cost groups (PMCGs) remain. Ward's minimum variance method (PROC CLUSTER, SAS Institute, 1999) is applied in order to derive the clusters, taking into account average differences between actual and expected costs on the hand and number of observations and variance on the other hand. The diagnostic referral codes 10 and 45 are removed from the analysis because of negative average differences between actual and expected costs. The decision to apply five clusters is based on the outcomes of the pseudo t2 statistic (not reported here). All five PMCGs should be added to the REF equation and tested for their contribution to the effectiveness of the risk-adjusted premium subsidies.

Medical device cost groups

Table 7.2 is based on a classification of medical devices that are delivered to Agis enrollees in 2001. This classification follows the classification of medical devices that is applied in CVZ (2002). The data are derived from the claims of medical devices in the Agis administration, except for the majority of claims for medical devices delivered by pharmacists because of feasibility problems.[125]

Table 7.2 shows the rank ordered categories of medical devices, after application of a rank ordering procedure that is analogous to the one applied with respect to the paramedic diagnostic referral codes. From the last table row it appears that 57.83 out of 1000 enrollees received some medical device in 2001, almost half of this subgroup received a dental prosthesis. Average actual costs lie € 1759 above the level of expenses that might be expected given the age, sex, and PCG composition of this subgroup of enrollees. It follows that 94.2% of the Agis enrollees form the subgroup for which no sold or hired medical device is found in the 2001 Agis claims data. The deviation of average actual costs from expected costs turns out to be largest for the subgroup of insured using nutritional aids.

In this case no subsequent clustering of medical device categories is applied, as the number of categories is limited and only slightly larger than the number of PCGs or DCGs. Therefore, these rank-ordered categories of medical devices define the so-called Medical Device Cost Groups (MDCGs). All MDCGs should be tested

124. The REF weights are more stable the larger the number of insured people within a subgroup defined by a REF adjuster.

125. The calculations in this analysis are based on the medical device claims data for 81.9% of the Agis population of insured who used a medical device in 2001.

Table 7.2: Rank ordering of 15 categories of 2001 medical devices, enrollees belong to one category only (2002 costs, N = 1.5 million enrollees of whom 85,360 are using medical devices)

Category number	Description	Average actual – expected costs (pipy) [a]	Weighted prevalence per 1000 enrollees [b]	Medical device cost groups
14	Nutritional aids [c]	26377 *	0.18	15
1	Incontinence and stoma aids	7688 *	2.50	14
9	Respiratory devices	6587 *	1.06	13
11	Communication and alarm devices	3885 *	6.26	12
12	Mobility devices	3675 *	3.27	11
8	Medical treatment aids	2872 *	0.80	10
15	Contraceptives and wigs	2761 *	0.51	9
3	Orthesis and orthopaedic footwear	2470 *	5.71	8
13	Visual devices	1969 *	0.52	7
6	Home care devices	1797 *	2.50	6
10	Prosthesis	707 *	2.72	5
7	Elastic support stockings	466 *	3.91	4
5	Diabetes devices	289	1.72	3
2	Dental prosthesis	273 *	25.06	2
4	Auditive devices	99	1.14	1
All	---	1759 *	57.83	---

[a] For each subgroup of enrollees associated with a specific category of medical devices, the average difference is presented between their actual costs and costs that might be expected given their 2001 age, sex, and (rank ordered) PCG categorization. Expected costs are derived given the estimated coefficients from a linear regression of 2002 costs on 2001 age/sex and (rank ordered) PCG dummy variables, restricted to the subpopulation of enrollees for which no medical device is claimed in 2001. These calculations are performed on the claims data of the Agis population, i.e. not on those of the respondents to the Agis Health Survey 2001 alone. Regression weights are applied to correct for the number of months of membership in 2002, partial months being rounded upward.

[b] A member who used more than one medical device, is assigned to the subgroup for which the average difference between actual costs and expected costs is largest. Reported prevalences are weighted in order to correct for the number of months of membership in 2002, partial months being rounded upward.

[c] From Table 7.2 it appears that for the enrollees using nutritional aids actual costs minus expected costs equals 26,377 euro on average. This relatively high figure is based on only 264 enrollees and is accompanied by a standard error of 2,552 euro, whereas in all other cases the standard error is less than 552 euro (these figures are not tabulated).

* Average difference between actual and expected costs is statistically significant (two-sided t-test, $p <= 0.05$).

for their contribution as an additional REF adjuster to improve the effectiveness of the risk-adjusted premium subsidies.

Mental Pharmacy-based Cost Groups

In Table 7.3 a classification of drugs acting on the nervous system is presented along the lines of Van Vliet and Lamers (2000), based on the ATC code that is part of the pharmaceutical claims records (WHO 1999).

Table 7.3: Mental disorders identifiable by ATC codes of pharmaceutical drugs acting on the nervous system (Van Vliet and Lamers 2000)

Category number	Disorders	ATC [a]	Drug classes
1	Depression	N06A	Antidepressants
2	Manic-depressive psychosis	N05AN01	Lithium
3	(Other) Psychosis	N05A (excl. N05AN01)	Antipsychotics
4	Anxiety and nervous condition	N05B	Anxiolytics

[a] ATC = Anatomical Therapeutic Chemical classification index (WHO 1999).

Table 7.4 shows the rank ordered categories of 2001 pharmaceutical drugs acting on the nervous system that are used for more than half a year, after application of a rank ordering procedure that is analogous to the one applied with respect to the paramedic diagnostic referral codes. From the last table row it appears that 83.66 out of 1000 enrollees belong to a so-called mental pharmacy-based costs group (MPCGs). Average actual costs lie € 1177 above the level of expenses that might be expected given the age, sex, and PCG composition of this subgroup of enrollees. It follows that 91.6% of the Agis enrollees are part of the subgroup for which no pharmaceutical drugs acting on the nervous system are found in the Agis claims data of 2001. It is recommended to test all categories of pharmaceutical drugs acting on the nervous system for their contribution to the effectiveness of the risk-adjusted premium subsidies.

Table 7.4: Rank ordering of five categories of 2001 pharmaceutical drugs acting on the nervous system that are used for more than half a year, enrollees belong to one category only (2002 costs, N = 1.5 million enrollees) [a]

Category number	Disorders	Average actual – expected costs (pipy) [b]	Weighted prevalence per 1000 enrollees [c]	Mental pharmacy-based cost groups
3	(Other) Psychosis	1400 *	11.67	4
4	Anxiety and nervous condition	1277 *	47.63	3
1	Depression	882 *	23.72	2
2	Manic-depressive psychosis	628 *	0.65	1
All	---	1177 *	83.66	---

[a] A member taking drugs that belong to more than one category of drugs, is classified in the subgroup for which the average difference between actual costs and expected costs is largest.

[b] For each subgroup of enrollees with a specific disorder, the average difference is presented between their actual costs and costs that might be expected given their 2001 age, sex, and (rank ordered) PCG categorization. Expected costs are derived given the estimated coefficients from a linear regression of 2002 costs on 2001 age/sex and (rank ordered) PCG dummy variables, restricted to the subpopulation of enrollees for which no drugs taken longer than 180 days are claimed in 2001. These calculations are performed on the claims data of the Agis population, i.e. not on those of the respondents to the Agis Health Survey 2001 alone. Enrollees are assigned to a category of drugs if the daily defined doses of the corresponding drugs exceeds 180 (WHO 1999, and Van de Ven, Van Vliet, and Lamers 2004). Regression weights are applied to correct for the number of months of membership in 2002, partial months being rounded upward.

[c] Reported prevalences are weighted in order to correct for the number of months of membership in 2002, partial months being rounded upward.

* Average difference between actual and expected costs is statistically significant (two-sided t-test, $p <= 0.05$).

Paramedic, Medical Device and Mental Pharmacy-based Cost Groups

Table 7.5 shows the REF weights that follow from equation (2.2) if the PMCGs, MDCGs and MPCGs risk adjusters are added to the REF equation. More specifically, the subgroups defined by these risk adjusters are:

1. Five subgroups of enrollees classified by diagnostic referral codes with a paramedic chronic indication (see Table 7.1); [126]
2. Fifteen subgroups of enrollees to whom medical devices are provided (see Table 7.2);
3. Four subgroups of enrollees to whom pharmaceutical drugs acting on the nervous system are provided for more than 180 days (see Table 7.4).

The old REF weights are copied from Table 6.13. Except for the subgroup of females between 15 and 24 years of age, the new REF weights are reduced for the age/sex subgroups after the PMCGs, MDCGs and MPCGs are included as additional REF adjusters. The largest reduction can be observed in the subgroups of older enrollees, which might be expected because use of physiotherapy, medical devices and mental health care is more prevalent amongst the elderly. This cost variation is now captured by the new REF adjusters instead of age. The substantially lower estimated weight with respect to disabled enrollees is also in line with expectations. Furthermore, all regional REF weights are reduced relative to the reference category. The same observation holds for the PCGs and DCGs, especially the reductions in REF weights with respect to the subgroups of insured people with Parkinson (PCG06) and neuromuscular disorders (PCG10) are quite substantial.

126. Diagnostic referral codes are not available for all ZAO enrollees (see also footnote 1). Therefore, for four out of five PMCGs, prevalences are between 30% and 50% of the prevalences that hold for the non-ZAO enrollees. For the other PMCG, prevalences are comparable.

Table 7.5: Estimation results for the REF equation (2.1) with the old REF weights and the new REF weights after the PMCGs, MDCGs and MPCGs are added to this equation as risk adjusters.

REF adjusters	Old REF weights [a]	New REF weights [b]	Change in weights
Intercept	623	666	43
M 15-24 (reference category)	---	---	---
M 25-34	-209	-229	-20
M 35-44	-316	-355	-39
M 45-54	464	378	-86
M 55-64	380	332	-48
M 65-74	1493	1326	-166
M 75-84	2796	2522	-274
M >=85	1598	891	-707
F 15-24	-109	-103	5
F 25-34	345	329	-16
F 35-44	0	-72	-72
F 45-54	194	64	-130
F 55-64	274	134	-140
F 65-74	1003	760	-243
F 75-84	2289	1742	-547
F >=85	1662	852	-810
Disabled	1437	1170	-267
Employed (reference category)	---	---	---
Social welfare	211	115	-96
Unemployed	214	176	-37
Retired	341	342	0
Self-Employed	-197	-162	35
Region 1	457	385	-72
Region 2	262	188	-74
Region 3	138	99	-38
Region 4	239	194	-45
Region 5	37	-49	-87
Region 6	-121	-166	-45
Region 7	-31	-74	-42
Region 8	-164	-206	-42
Region 9	29	-16	-45
Region 10 (reference category)	---	---	---
No PCG (reference category)	---	---	---
PCG01	1883	1669	-214
PCG02	1803	1599	-203
PCG03	1098	1107	9

REF adjusters	Old REF weights [a]	New REF weights [b]	Change in weights
PCG04	2001	1742	-259
PCG05	3848	3381	-467
PCG06	3199	1492	-1707
PCG07	3366	3198	-169
PCG08	7791	7578	-213
PCG09	3823	3515	-308
PCG10	8030	5612	-2418
PCG11	11895	11954	59
PCG12	20748	20832	84
No DCG (reference category)	---	---	---
DCG01	1356	791	-564
DCG02	6319	6212	-108
DCG03	3565	3318	-247
DCG04	5591	4889	-701
DCG05	4262	3997	-265
DCG06	7820	6758	-1061
DCG07	6038	5285	-753
DCG08	8869	6824	-2046
DCG09	7983	6398	-1585
DCG10	18152	16925	-1227
DCG11	12626	11780	-846
DCG12	9050	8239	-811
DCG13	77982	77233	-749
No PMCG (reference category)	---	---	---
PMCG01	---	274	---
PMCG02	---	331	---
PMCG03	---	3286	---
PMCG04	---	1666	---
PMCG05	---	5123	---
No MDCG (reference category)	---	---	---
MDCG01	---	559	---
MDCG02	---	595	---
MDCG03	---	-474	---
MDCG04	---	199	---
MDCG05	---	165	---
MDCG06	---	635	---
MDCG07	---	-302	---
MDCG08	---	1303	---
MDCG09	---	2353	---
MDCG10	---	4678	---

REF adjusters	Old REF weights [a]	New REF weights [b]	Change in weights
MDCG11	---	4196	---
MDCG12	---	2551	---
MDCG13	---	8293	---
MDCG14	---	5903	---
MDCG15	---	5034	---
No MPCG (reference category)	---	---	---
MPCG01	---	-781	---
MPCG02	---	656	---
MPCG03	---	572	---
MPCG04	---	600	---
R^2_{ADJ}	17.93%	19.11%	

Note: PCGs, DCGs, PMCGs, MDCGs and MPCGs are rank ordered. See Tables 7.1, 7.2 and 7.4 for the classification of the PMCGs, MDCGs and MPCGs, respectively. PCG01=Asthma/COPD, PCG02=Epilepsy, PCG03=Crohn/Colitis Ulcerosa, PCG04=Cardiac disease, PCG05=Rheumatism, PCG06=Parkinson, PCG07=Diabetes (Type I), PCG08=Transplantation, PCG09=Cystic fibrosis, PCG10=Neuromuscular disorder, PCG11=HIV/Aids, PCG12=Renal disease/ESRD.

[a] These estimates result after the estimation of REF equation (2.1).

[b] These estimates result after the estimation of REF equation (2.1) with the PMCGs, MDCGs and MPCGs added as new REF adjusters.

These PCGs are based on prescribed drugs acting on the nervous system and their effects are probably taken over by the MPCGsin particular.

The estimated REF weights associated with the PMCGs, MDCGs and MPCGs follow a non-monotonous pattern, sometimes even a negative weight holds. Negative weights might be caused by a low prevalence of enrollees assigned to the respective subgroups of enrollees using diabetes devices (MDCG03), visual devices (MDCG07) and pharmaceutical drugs against manic-depressive psychosis (MPCG01). The adjusted R^2 increases from 17.93% to 19.11% by adding the aforementioned new REF adjusters.

Table 7.6 shows that, after addition of the new set of REF adjusters to the REF equation, the gap between REF predicted costs and normative costs substantially reduces for nearly all subgroups defined by the S-type adjusters. The weighted average of the absolute values of the reported deviations from normative costs following equation (2.6) equals 157, if all subgroups defined by the S-type adjusters inclusive those defined by age, gender, the PCGS and DCGs are taken into account.[127] Therefore, if the risk-adjusted premium subsidies would be based on

127. See the Appendix A7.1 for the deviations of REF predicted costs from normative costs with respect to age, gender, the PCGs and DCGs.

Table 7.6: REF predicted costs compared to normative costs 2002 given the old and new REF weights (see Table 7.5), for survey respondents grouped by the S-type adjusters from the normative equation [a]

		REF predicted costs – normative costs (pipy)		REF predicted costs / normative costs	
Subgroups	**Size of subgroup**	**Old REF weights**	**New REF weights**	**Old REF weights**	**New REF weights**
Q1 PF scores	25.0%	-752 *	-630 *	0.799 *	0.832 *
Q1 RP scores	25.0%	-621 *	-485 *	0.823 *	0.862 *
Q1 BP scores	25.0%	-628 *	-497 *	0.800 *	0.842 *
Q1 GH scores	25.0%	-599 *	-482 *	0.829 *	0.863 *
Q1 VT scores	25.0%	-419 *	-311 *	0.853 *	0.891 *
Q1 SF scores	25.0%	-514 *	-400 *	0.826 *	0.864 *
Q1 RE scores	25.0%	-318 *	-232 *	0.882 *	0.914 *
Q1 MH scores	25.0%	-221 *	-134 *	0.905 *	0.942 *
Q2 PF scores	25.0%	124 *	85 *	1.073 *	1.050 *
Q2 RP scores	25.0%	77 *	46	1.050 *	1.030
Q2 BP scores	25.0%	40	12	1.024	1.007
Q2 GH scores	25.0%	46 **	15	1.028 **	1.009
Q2 VT scores	25.0%	-3	-13	0.999	0.992
Q2 SF scores	25.0%	1	-11	1.000	0.994
Q2 RE scores	25.0%	103 *	78 *	1.070 *	1.053 *
Q2 MH scores	25.0%	4	-7	1.002	0.996
Q3 PF scores	25.0%	292 *	253 *	1.308 *	1.267 *
Q3 RP scores	25.0%	288 *	222 *	1.272 *	1.211 *
Q3 BP scores	25.0%	252 *	206 *	1.221 *	1.181 *
Q3 GH scores	25.0%	225 *	178 *	1.200 *	1.159 *
Q3 VT scores	25.0%	163 *	128 *	1.126 *	1.099 *
Q3 SF scores	25.0%	250 *	190 *	1.211 *	1.160 *
Q3 RE scores	25.0%	118 *	77 **	1.074 *	1.049 **
Q3 MH scores	25.0%	93 *	58 **	1.065 *	1.040 **
Q4 PF scores	25.0%	336 *	292 *	1.541 *	1.470 *
Q4 RP scores	25.0%	256 *	216 *	1.278 *	1.234 *
Q4 BP scores	25.0%	336 *	279 *	1.319 *	1.264 *
Q4 GH scores	25.0%	328 *	289 *	1.457 *	1.402 *
Q4 VT scores	25.0%	259 *	196 *	1.232 *	1.176 *
Q4 SF scores	25.0%	263 *	221 *	1.254 *	1.213 *
Q4 RE scores	25.0%	97 *	76 *	1.076 *	1.060 *
Q4 MH scores	25.0%	123 *	83 *	1.087 *	1.059 *
Number of self-reported OECD limitations [b]					
0	77.8%	204 *	165 *	1.168 *	1.136 *
1	10.3%	-370 *	-345 *	0.873 *	0.881 *
2	4.9%	-916 *	-767 *	0.768 *	0.806 *

Subgroups	Size of subgroup	REF predicted costs – normative costs (pipy)		REF predicted costs / normative costs	
		Old REF weights	New REF weights	Old REF weights	New REF weights
3+	5.2%	-1445 *	-1066 *	0.726 *	0.798 *
Imputed	1.8%	-13	17	0.994	1.008
Number of self-reported chronic conditions [b]					
0	90.4%	93 *	82 *	1.066 *	1.058 *
1	8.1%	-778 *	-703 *	0.831 *	0.847 *
2	1.3%	-1268 *	-978 *	0.806 *	0.850 *
3+	0.2%	-2511 *	-2075 *	0.726 *	0.774 *
Total	100.0%	0	0	1.000	1.000

*,** Difference between average predicted and normative costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

[a] The weighted average of the absolute values of the differences between REF predicted costs and normative costs for the tabulated subgroups equals 199. If the subgroups defined by the S-type adjusters age, gender, PCGs and DCGs are also taken into account (see the Appendix A7.1) then this figure equals 157.

[b] The sizes of these subgroups sum up to 100%.

this new set of REF adjusters, the policy goals of the Dutch government can be achieved up to an extent of (1-157/687) x 100% = 77.1% in this way.[128] Remember that this figure equals 71.2% for the original 2004 Dutch REF equation with the old set of REF adjusters (see Table 6.6). The PMCGs, MDCGs and MPCGs are therefore good candidates to be added as new risk adjusters to the 2004 Dutch REF equation.

In this section, a normative test of potential new REF adjusters has been performed. If PMCGs, MDCGs and MPCGs are added to the 2004 Dutch REF equation, it shows that these reduce the gap between REF predicted costs and normative costs substantially. Nevertheless, some part of the gap with normative costs remains. In Section 7.2 risk sharing is tested as a supplement to incomplete and/or imperfect REF adjusters.

7.2 Ex-post risk sharing as a supplement to the REF equation

In this section, risk sharing analogous to the 2004 Dutch risk sharing scheme is introduced, as an alternative way to improve the effectiveness of the risk-adjusted premium subsidies as tested in Chapter Six. Risk sharing means that deviations of

128. The figure of 687 can be found in Table A6.1, footnote a.

actual costs from REF predicted costs are retrospectively shared between insurers and the Risk Equalization Fund to some predetermined extent. The ex-post risk-sharing scheme applied in this study consists of 100% reimbursement of the production-independent hospital costs, and reimbursement of 90% of the production-dependent hospital costs, medical specialist costs and outpatient costs above a threshold of € 12,500 (i.e. a combination of "outlier risk sharing" and "proportional risk sharing").[129] A hospital specific separation of hospital costs into production-independent costs on the one hand and production-dependent hospital costs and medical specialist costs on the other hand is applied in the Dutch REF model since 2002.[130]

In order to apply full reimbursement to the so-called "fixed" part of hospital costs, total hospital costs must be split into a fixed part of production independent hospital costs on the one hand and production dependent hospital costs on the other hand. In the 2004 Dutch REF model formally 95% of the "fixed" hospital costs are reimbursed retrospectively, whereas the remaining 5% are reimbursed prospectively in the form of per capita payments. These per capita payments are not risk rated, however. Under the assumption that these per capita payments sum up to the total amount of production-independent hospital costs in the current research sample, this scheme boils down to full retrospective reimbursement.

Furthermore, the 2004 Dutch REF model comprises two additional proportional risk sharing schemes with respect to the production-dependent and medical specialist costs, which – combined – lead to risk sharing between insurers and the REF of about 55%. In this study, these additional proportional risk sharing schemes are not included in order not to complicate matters unnecessarily. The conclusions drawn are not altered by this simplification.

Table 7.7 shows that REF predicted costs are closer to normative costs for the 2004 Dutch REF equation with risk sharing than without risk sharing. The gap appears to reduce for nearly all subgroups defined by the S-type adjusters.[131] The weighted average of the absolute values of the reported deviations from normative costs following equation (2.6) equals 141 after risk sharing is applied, such that the policy goals of the Dutch government can be achieved up to an extent of

129. Along the lines of Van Vliet (2005), in the calculations a threshold of € 10.732 is applied instead of € 12.500 in order to compensate for yearly price inflation of close to 8% between 2002 (= data year) and 2004 (= year under study).

130. See also Table A1.3 in Chapter One. The CVZ (2003) average ex-post variable hospital tariffs 2002 are applied in this study in order to split hospital specific per diem tariffs into a production-independent part and a production-dependent and medical specialists part.

131. See the Appendix A7.2 for the results with respect to age, gender, the PCGS and DCGs.

Table 7.7: REF predicted costs with and without risk sharing compared to normative costs 2002, for survey respondents grouped by the S-type adjusters from the normative equation [a]

		REF predicted costs – normative costs (pipy)		REF predicted costs / normative costs	
Subgroups	**Size of subgroup**	**Without risk sharing**	**Including risk sharing**	**Without risk sharing**	**Including risk sharing**
Q1 PF scores	25.0%	-752 *	-536 *	0.799 *	0.857 *
Q1 RP scores	25.0%	-621 *	-442 *	0.823 *	0.874 *
Q1 BP scores	25.0%	-628 *	-407 *	0.800 *	0.870 *
Q1 GH scores	25.0%	-599 *	-376 *	0.829 *	0.893 *
Q1 VT scores	25.0%	-419 *	-280 *	0.853 *	0.902 *
Q1 SF scores	25.0%	-514 *	-327 *	0.826 *	0.889 *
Q1 RE scores	25.0%	-318 *	-220 *	0.882 *	0.918 *
Q1 MH scores	25.0%	-221 *	-150 *	0.905 *	0.936 *
Q2 PF scores	25.0%	124 *	17	1.073 *	1.010
Q2 RP scores	25.0%	77 *	85 *	1.050 *	1.055 *
Q2 BP scores	25.0%	40	-70 **	1.024	0.958 **
Q2 GH scores	25.0%	46 **	-64 *	1.028 **	0.962 *
Q2 VT scores	25.0%	-3	-9	0.999	0.995
Q2 SF scores	25.0%	1	-76 **	1.000	0.959 **
Q2 RE scores	25.0%	103 *	88 *	1.070 *	1.060 *
Q2 MH scores	25.0%	4	-48	1.002	0.974
Q3 PF scores	25.0%	292 *	237 *	1.308 *	1.250 *
Q3 RP scores	25.0%	288 *	189 *	1.272 *	1.179 *
Q3 BP scores	25.0%	252 *	207 *	1.221 *	1.181 *
Q3 GH scores	25.0%	225 *	137 *	1.200 *	1.122 *
Q3 VT scores	25.0%	163 *	63 *	1.126 *	1.049 *
Q3 SF scores	25.0%	250 *	233 *	1.211 *	1.196 *
Q3 RE scores	25.0%	118 *	100 *	1.074 *	1.063 *
Q3 MH scores	25.0%	93 *	56 **	1.065 *	1.039 **
Q4 PF scores	25.0%	336 *	281 *	1.541 *	1.453 *
Q4 RP scores	25.0%	256 *	169 *	1.278 *	1.183 *
Q4 BP scores	25.0%	336 *	271 *	1.319 *	1.257 *
Q4 GH scores	25.0%	328 *	303 *	1.457 *	1.422 *
Q4 VT scores	25.0%	259 *	225 *	1.232 *	1.202 *
Q4 SF scores	25.0%	263 *	169 *	1.254 *	1.163 *
Q4 RE scores	25.0%	97 *	32	1.076 *	1.025
Q4 MH scores	25.0%	123 *	143 *	1.087 *	1.101 *
Number of self-reported OECD limitations [b]					
0	77.8%	204 *	123 *	1.168 *	1.101 *
1	10.3%	-370 *	-231 *	0.873 *	0.921 *
2	4.9%	-916 *	-568 *	0.768 *	0.856 *

Subgroups	Size of subgroup	REF predicted costs – normative costs (pipy)		REF predicted costs / normative costs	
		Without risk sharing	Including risk sharing	Without risk sharing	Including risk sharing
3+	5.2%	-1445 *	-897 *	0.726 *	0.830 *
Imputed	1.8%	-13	183	0.994	1.090
Number of self-reported chronic conditions [b]					
0	90.4%	93 *	63 *	1.066 *	1.044 *
1	8.1%	-778 *	-511 *	0.831 *	0.889 *
2	1.3%	-1268 *	-1024 *	0.806 *	0.843 *
3+	0.2%	-2511 *	-1004 *	0.726 *	0.891 *
Total	100.0%	0	0	1.000	1.000

*,** Difference between average predicted and normative costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

[a] The weighted average of the absolute values of the differences between REF predicted costs and normative costs for the tabulated subgroups equals 180. If the subgroups defined by the S-type adjusters age, gender, PCGs and DCGs are also taken into account (see the Appendix A7.2) then this figure equals 141.

[b] The sizes of these subgroups sum up to 100%.

(1-141/687) x 100% = 79.5%.[132] Note that this performance figure is even larger than the 77.1% that emerges after addition of PMCGs, MDCGs and MPCGs to the REF equation (see Section 7.1). This observation suggests that ex-post risk sharing may remain an important supplement to imperfect risk adjusters, even after the set of REF adjusters in the Dutch REF equation is improved.

In this section, a normative test of risk sharing as a supplement to incomplete and/or imperfect REF adjusters has been performed. In case of a risk sharing scheme that closely resembles the 2004 Dutch risk sharing scheme, it shows that the gap between REF predicted costs and normative costs is substantially reduced by ex-post risk sharing. Ex-post risk sharing remains an important supplement to imperfect risk adjusters, even after the set of risk adjusters proposed in Section 7.1 are added to the 2004 Dutch REF equation.

7.3 The functional form and error distribution of the REF equation

Nearly all premium models in the insurance literature are based on a multiplicative specification. However, in the risk equalization literature the specification of

132. The figure of 687 can be found in Table A6.1, footnote a.

the REF equation is often additive. In this section, the additive specification of the 2004 Dutch REF equation is tested against an alternative specification that assumes multiplicative effects of the REF adjusters on costs by applying a log link function in a GLM framework. This GLM framework also allows to tackle skewness and potential heteroscedasticity by choosing an appropriate specification of the error distribution and variance function. Given the present research sample, REF predicted costs are estimated under this alternative specification and compared to the original REF predicted costs.

Before 2002 the Dutch REF equation was multiplicative in its REF adjusters: insurance eligibility (interacted with age) and region were included as multiplicative REF adjusters with respect to the base REF predicted costs, which depended on age and gender.[133] Given the base REF predicted costs, a stepwise and cell-based estimation procedure was followed in order to find the multiplicative factors for insurance eligibility and region. A multiplicative specification takes care of situations, for example, in which it is expected that elderly enrollees being disabled and living in highly urbanized regions might induce higher costs than what is expected if the average weights of these REF adjusters are merely added together. As only demographic variables were included in the before 2002 Dutch REF scheme which are rather crude and indirect indicators of health status, a multiplicative specification may do a better job in capturing bad health situations that can be seen as a cumulation of adverse background characteristics.

Since the introduction of PCGs in 2002 the REF formula is additive instead of multiplicative, because excessive values of REF predicted costs would arise for specific subgroups if a multiplicative specification would be continued. An additional advantage of an additive specification is that REF predicted costs calculated at the individual level of the insured sum up to the same amount of money as when calculated at the aggregated level of an insurer.

In 2002 not only the multiplicative specification was replaced by an additive specification of the REF equation, but also the stepwise and cell-based approach to estimate the REF weights was replaced by the method of least-squares such that the weights of all the REF adjusters are estimated simultaneously. This means that instead of a stepwise minimization of the differences between REF predicted costs and observed costs for each REF adjuster separately, the sum of squared differences for each combination of the REF adjusters is minimized in one step. If the parameters are estimated by the method of least-squares, then the statistical tests about the estimators are derived under the assumption of normally dis-

133. See Table A1.2 in Chapter One.

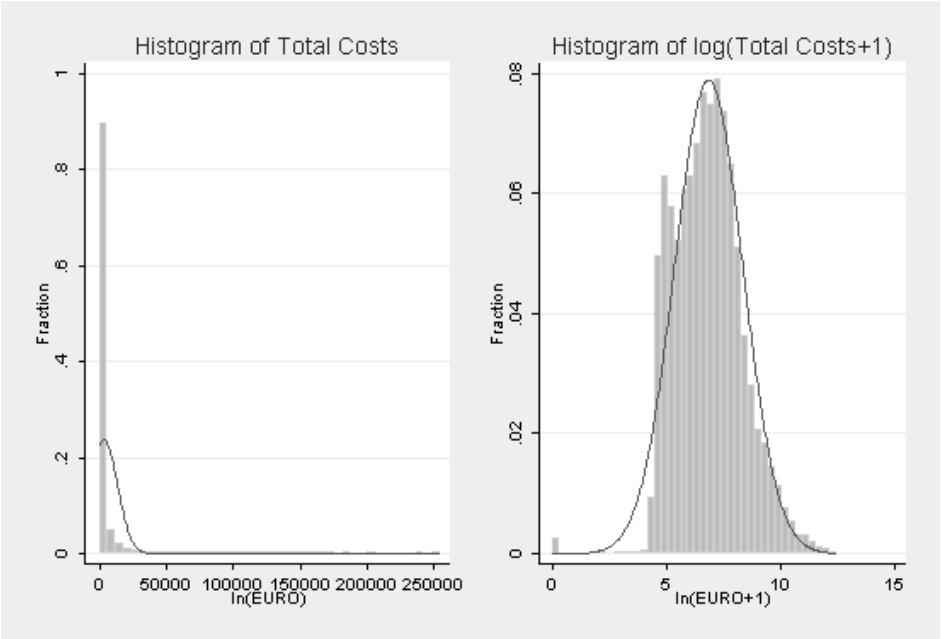


Figure 7.1: Histogram of 2002 health care costs and a logarithmic transformation of 1 EURO plus these costs.

tributed error terms with constant error variances (i.e. homoscedasticity) across subgroups as defined by the REF adjusters.

From the left panel of Figure 7.1 it is obvious that 2002 health care costs in our sample are heavily skewed instead of normally distributed. Distributional skewness of costs is typical in health care for several reasons: of course observed costs are non-negative, often only a small portion of patients incurs very high health care costs and/or often there is a substantial portion of those who are insured with zero costs.[134] The skewed distribution of health care costs does not pose a problem in the context of risk equalization in Dutch practice, because the parameters of the REF model are based on the total population of about 10 million insured people in this case.[135] On the other hand, skewness may pose problems in this study given the limited size of the research sample. The standard procedure to take account of skewness of health care data is to choose an appropriate error distribution function, for example, the lognormal distribution. Therefore, in this

134. In the present research sample there is a substantial portion of those who are insured with positive instead of zero costs, as every insured should be registered with a GP and thus these subscription costs are claimed for all insured. This does not hold for 0.12% of all enrollees who are not registered with a GP, for example because they are staying at a nursing home.

135. Under the Health Insurance Act 2006 the insured population consists of all 16 million Dutch inhabitants.

section, distributions are applied that explicitly take account of this skewness when modeling health care costs for the research sample.

In addition to skewness of health care costs, heteroscedasticity between subgroups of insured people is also often observed. Heteroscedasticity means that the error variances are not constant across subgroups. However, the possibility of heteroscedasticity is not a relevant problem for the purpose of risk equalization models, neither in this study, nor in practice. The reason for this is that the OLS estimate of the REF weights – which is the statistic of interest in this case – will remain unbiased and consistent under heteroscedasticity. Heteroscedasticity should only raise efficiency concerns as a consequence of the limited research sample used in this study: these estimates are no longer best linear unbiased (i.e. smallest sampling variance in the class of all linear unbiased estimators) or asymptotically efficient (i.e. sampling distribution collapses most quickly on to the regression parameters being estimated in the class of all consistent estimators).[136] Note that these efficiency concerns can safely be ignored in the context of risk equalization models in practice, because the estimation of the model parameters is based on the claims data of the total population of about 10 million Dutch insured people in 2004.

Multiplicativity, skewness and potential heteroscedasticity must be dealt with by choosing an appropriate functional relationship between costs and the REF adjusters. In order to determine which transformation fits the data best, a Box-Cox (1964) transformation procedure can be applied. This transformation is given by rewriting REF equation (2.1) as follows:

$$(7.1)\quad \frac{Y_t^{\lambda}-1}{\lambda}=\alpha_0+\sum_{j=1}^{J}\alpha_j X_{j,t-1}+\epsilon_t$$

where Y_t are 2002 costs and λ is the transformation parameter to be estimated. In the limiting case of λ being equal to zero, by L'Hôpital's rule, the REF equation (2.1) boils down to the following log link relationship:

$$(7.2)\quad \log(Y_t)=\alpha_0+\sum_{j=1}^{J}\alpha_j * X_{j,t-1}+\epsilon_t$$

A log transformation captures a multiplicative relationship between the REF adjusters at the untransformed scale, which can be seen by exponentiating both

136. As a consequence, standard confidence intervals for and hypotheses testing procedures about regression parameters can no longer be relied upon (Thomas 1985). The standard procedure to solve for heteroscedasticity is to assume a specific form of error variance function and apply generalized least squares (GLS) to the REF equation.

Table 7.8: Box-Cox test of null hypothesis $H_0 : \lambda = \lambda\hat{}$[a]

$\lambda\hat{}$	Type of link function	Likelihood ratio test χ2(54)
-1	Reciprocal link	1.0e+05 * (p<0.001)
0	Log link	645.60 * (p<0.001)
1	Linear link	81214.57 * (p<0.001)

* The true coefficient differs statistically significant from the hypothesized value (two-sided t-test, p <= 0.05).

[a] The Box-Cox test is performed without applying sampling weights to the individual observations.

sides of equation 7.2. Furthermore, under the assumption that the residuals ε_t are normally distributed, a so-called lognormal model for Y_t at the untranformed scale is defined in this way. Lognormal models are often applied to deal with distributions being skewed to the right. Other common transformations are also included as special cases of the Box-Cox transformation procedure, such as the reciprocal transformation (λ = -1) and the square root transformation (λ = 0.5).

A maximum likelihood optimization procedure is applied in order to find the λ parameter value that fits the present data best. The estimation results of this Box-Cox test (without applying population weights) gives an estimated λ value of -.0134 with a 95% confidence interval of (-.0196, -.0071). Although the hypothesis of a zero λ is rejected, from the results of a likelihood ratio test of λ values of -1 and 1 it appears that the logarithmic transformation of 2002 costs is the most appropriate parametric transformation for this study sample.

Given a logarithmic transformation procedure applied to the left-hand side of REF equation (2.1), explicitly described by equation (7.2), a retransformation is needed to obtain REF predicted costs. However, such a retransformation procedure is not straightforward. This can be seen by exponentiating both sides of equation (7.2) which gives

$$(7.3)\quad Y_t = \exp(\alpha_0) * \prod_{j=1}^{J} \exp(\alpha_j * X_{j,t-1}) * \exp(\epsilon_t)$$

and noticing that the error term has a non-constant influence on REF predicted costs $E[Y_t|X_{j,t-1}, j=1,...,J]$ if the error term ε_t is heteroscedastic in some $X_{j,t-1}$ or any other non-included variable, where E[.] is the expectation operator.

Suppose that the error term is normally distributed and the parametric distribution is known in advance, for example, $\varepsilon_t \sim N(0, \sigma^2(X_{j,t-1}))$ where $\sigma^2(X_{j,t-1})$ denotes that the error variance is a function of the j^{th} REF adjuster. Then REF predicted costs $\hat{Y}_{t+1}=E[Y_t|X_{j,t-1}, j=1,\ldots,J]$ can be written as

$$(7.4)\quad \hat{Y}_t = \exp(\hat{\alpha}_0) * \prod_{j=1}^{J} \exp(\hat{\alpha}_j * X_{j,t-1}) * \exp(\frac{1}{2}\hat{\sigma}^2)$$

In general, REF predicted costs can not simply be derived by exponentiating predictions of logarithmic costs, in case of heteroscedastic residuals at the transformed scale in equation (7.2). In order to derive REF predicted costs, the exponentiated predictions have to be multiplied by a factor that is based on the heteroscedastic error variance $\sigma^2(X_{j,t-1})$, in this case as described by the last factor in equation (7.4). Heteroscedasticity at the transformed scale leads to biased estimators on the untransformed scale if not properly retransformed. Therefore, in contrast to heteroscedasticity on the untransformed scale, heteroscedasticity can not be ignored if a transformed version of the REF equation is estimated.

If the error term is not normally distributed but the parametric distribution is known in advance, then the expectation of $\exp(\varepsilon_t)$ can be derived analytically under heteroscedasticity. If the parametric distribution is not known in advance, then a consistent estimate of the expectation of $\exp(\varepsilon_t)$ may be obtained by applying a so-called non-parametric smearing factor (Duan 1983). This smearing factor is equal to the average of exponentiated estimates of the transformed residuals and removes the bias in REF predicted costs.

As an alternative to applying a direct transformation to the 2002 costs, an appropriate so-called Generalized Linear Model (GLM) can be chosen such that the retransformation problem does not arise. The GLM framework is characterized by the following two equations:

$$(7.5)\quad Y_t = \hat{Y}_t + \hat{\epsilon}_t$$

and

$$(7.6)\quad g(\hat{Y}_t) = \hat{\alpha}_0 + \sum_{j=1}^{J} \hat{\alpha}_j * X_{j,t-1}$$

where $g(.)$ is called the link function that defines the type of relationship between expected 2002 costs and the REF adjusters. By choosing a specific link function and error distribution, a wide variety of models can be represented. For example, in case of an identity link and a normal distribution the standard linear regression model is obtained.

Note that in the GLM framework the link function does not introduce retransformation problems. REF predicted costs $E[Y_t]$ can simply be derived by the retransformation of $g^{-1}(g(E[Y_t]))$ in the GLM framework, provided *g(.)* is monotone and differentiable. In case of a log link there holds

$$(7.7)\quad \hat{Y}_t = \exp(\hat{\alpha}_0) * \prod_{j=1}^{J} \exp(\hat{\alpha}_j * X_{j,t-1})$$

without the need for a multiplication factor that corrects for heteroscedasticity in the residuals at the transformed scale.

In addition to the choice of a log link, an error distribution has to be chosen in the GLM framework that describes the mean-variance relationships. In general, the relationship between mean and variance in the GLM framework can be written mathematically as

$$(7.8)\quad \mathrm{Var}(\hat{Y}_t) = \phi * [\hat{Y}_t]^{\gamma}$$

where γ must be finite and non-negative. The value of γ determines which specific distribution fits the data best. If γ=0 then the usual non-linear least-squares estimator is obtained in which case variance is unrelated to the mean; if γ=1 then the Poisson-like class holds and variance equals the mean. The γ=2 assumption holds in case of a gamma distribution, the homoscedastic log normal, the Weibull and the Chi-Square, and variance exceeds the mean. The inverse Gaussian or Wald distribution is obtained in case of γ=3 (Blough, Madden and Hornbrook 1999, Manning and Mullahy 2001).

Taking natural logarithms of both sides of equation (7.8) gives

$$(7.9)\quad \ln(\mathrm{Var}(\hat{Y}_t)) = \ln(\phi) + \gamma * \hat{Y}_t$$

Adding $\ln(\hat{\varepsilon}_t^2)$ to both sides of equation (7.9) and rewriting then gives

$$(7.10)\quad \ln(\hat{\epsilon}_t^2) = \phi' + \gamma * \hat{Y}_t + \nu_t$$

where $\phi'=\ln(\phi)$ and $\nu_t=\ln(\hat{\varepsilon}_t^2/\mathrm{Var}(\hat{Y}_t))$. Equation (7.10) forms the basis for a modified Park test of the type of heteroscedasticity as described by equation (7.8), i.e. the logarithm of squared untransformed scale residuals $\hat{\varepsilon}_t^2$ is robustly regressed on the logarithm of REF predicted costs $\hat{Y}_t$ and a constant.[137] The estimated coefficient $\hat{\gamma}$ then determines the relationship between the variance and the mean as described in equation (7.8) and thus identifies the appropriate family (distribution) function among the GLM alternatives (Manning and Mullahy 2001; Manning, Basu

137. Note that $\hat{\varepsilon}_t^2$ and the $\hat{\alpha}_j$, $j=0,1,\ldots,J$ parameters result after estimation of equation (7.6) under the assumption of a log link, either by applying the OLS estimation technique or by applying the GLM estimation technique under the additional assumption of a gamma distribution (Manning and Mullahy 2001).

Table 7.9: Modified Park test to determine choice of preferred type of distribution in a GLM framework.

Hypothesis about estimated coefficient γ	Type of model distribution	F(1, 18615), given GLM estimates [a]	F(1, 18615), given OLS estimates [b]
0	Gaussian	9868.84 * (p<0.0001)	6826.64 * (p<0.0001)
1	Poisson	2529.77 * (p<0.0001)	1795.87 * (p<0.0001)
2	Gamma	1.57 (p=0.211)	4.55 * (0.033)
3	Wald or inverse Gaussian	2284.23 * (p<0.0001)	1452.65 * (p<0.0001)

* The true coefficient differs statistically significant from the hypothesized value (two-sided t-test, p <= 0.05).

[a] The $\hat{\varepsilon}_t^2$ and $\hat{\alpha}_j$, j=0,1,…,J parameters in the modified Park test equation (7.10) result from GLM estimation of equation (7.6), under the assumption of a log link and a gamma distribution.

[b] The $\hat{\varepsilon}_t^2$ and $\hat{\alpha}_j$, j=0,1,…,J parameters in the modified Park test equation (7.10) result from OLS estimation of equation (7.6), under the assumption of a log link.

and Mullahy 2005). Note that the modified Park test requires a correct specification of the link function.

The results of the modified Park test are presented in Table 7.9.[138] Under the assumption of a log link, only the gamma distribution seems to be able to describe the relationship between squared residuals and REF predicted costs: a γ value of two can not be rejected given a 95% level of confidence if the modified Park test equation (7.10) is estimated by GLM under the assumption of a gamma distribution. If equation (7.10) is estimated by OLS, this conclusion holds under the assumption of a 1% level of significance.

For the present study sample, $\hat{\gamma}$ equals 2.026 (robust standard error: 0.020). In Figure 7.2 the residual variances $\ln(\hat{\varepsilon}_t^2)$ are graphed against REF predicted costs for each enrollee, confirming that a slope equal to two seems to be a good approximation of the linear relationship between these variables.

It has become common practice to assume a gamma distribution and a log link when dealing with health care expenditures. Basu, Manning and Mullahy (2004) compare bias in the parameter estimates when applying OLS on the logarithmic of observed costs, a gamma regression model with a log link (i.e. a multiplicative error term), and Cox proportional hazards regression. No single alternative appears best under all of the conditions examined by them (see also Manning and Mullahy

138. Results presented in this section are calculated in Stata version 9.2 (StataCorp 2006). Prof. Basu is acknowledged for kindly providing the basic Stata do file. For the purpose of this study, the statistical procedures contained in this file were extended by allowing for the application of sampling weights. This version of the computer code is available from the author upon request.

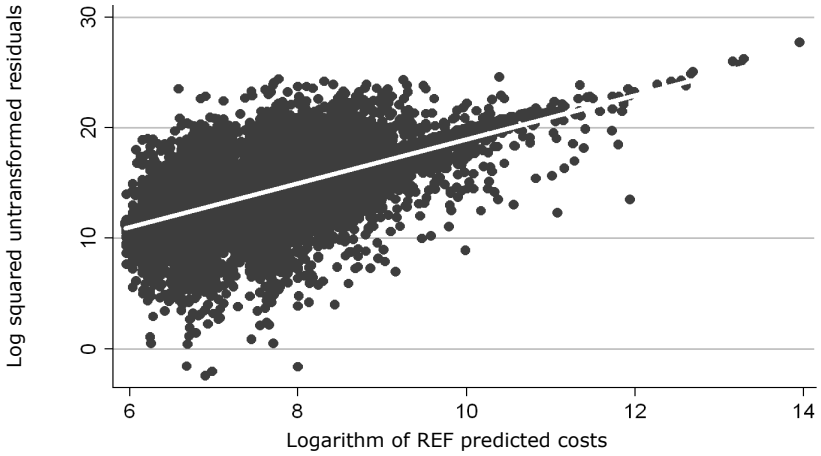


Figure 7.2: Logarithm of squared untransformed residuals $\ln(\hat{\varepsilon}_t^2)$ versus logarithm of REF predicted costs $\ln(\hat{Y}_t)$, after GLM estimation of equation (7.10) with log link and gamma distribution

2001). However, they find that the gamma regression model with a log link is most robust to the true underlying population distributions of strictly positive values of hospital claims. Note that besides the fact that a gamma distribution is able to describe a specific form of heteroscedasticity in the error variances, i.e. error variances are assumed to be proportional to the square of the error mean, they can also describe different shapes, e.g. steadily declining from zero, or bell-shaped but skewed to the left.

In Manning, Basu and Mullahy (2005) a three parameter generalized Gamma (GGM) distribution is employed that includes several of the standard alternative models as special cases: OLS with a normal error, OLS for the log-normal transformation, the standard Gamma with a log link, and the Weibull. From a simulation study, Manning, Basu and Mullahy (2005) conclude that the nested GGM modeling strategy may provide a more robust alternative estimator to the standard alternatives mentioned above.

Table 7.10 shows the testing results from the GGM estimations of the aforementioned models including the GGM model under the assumption of homoscedastic and heteroscedastic errors, respectively.[139] Four tests for goodness of fit are presented in order to be able to choose the optimal set of three parameters and thereby the preferred model structure (see also Manning, Basu and Mullahy

139. Prof. W.G. Manning is acknowledged for kindly providing the basic Stata ado file and instructions for GGM estimation upon request.

Table 7.10: Goodness of fit on the untransformed scale of 2002 health care costs + 1 euro, given the REF adjusters as explanatory variables.

Estimator	Hosmer-Lemeshow test	Pregibon Link test	Ramsey RESET test	Pearson correlation test	Average residual
OLS for y	2.21 * (p=0.0146)	0.25 (p=0.6184)	2.02 (p=0.1081)	0.00 [a] (p=1.0000)	0.00
OLS for ln(y)	278.88 * (p<0.0001)	20.30 * (p<0.0001)	45.79 * (p<0.0001)	0.28 * (p<0.0001)	1754.29
Gamma regression	2.17 * (p=0.0166)	17.29 * (p<0.0001)	84.93 * (p<0.0001)	-0.46 * (p<0.0001)	-129.99
Weibull regression	2.81 * (p<0.0001)	17.91 * (p<0.0001)	33.85 * (p<0.0001)	-0.52 * (p<0.0001)	-13.08
GGM [b]	97.90 * (p<0.0001)	11.99 * (p=0.0005)	30.56 * (p<0.0001)	-0.28 * (p<0.0001)	982.23
GGM – heteroscedastic errors [b,c]	97.31* (p<0.0001)	12.70 * (p=0.0004)	34.86 * (p<0.0001)	-0.28 * (p<0.0001)	970.02

Tests for identifying distributions		χ^2 statistic	Degrees of freedom	p-value
Standard gamma	$\kappa = \sigma$	2341.22	1	<0.0001
Log-normal	$\kappa = 0$	63.59	1	<0.0001
Weibull	$\kappa = 1$	1142.52	1	<0.0001
Exponential	$\kappa = \sigma = 1$	4314.60	2	<0.0001

* The test statistic differs statistically significant from zero (two-sided t-test, p <= 0.05).
[a] This follows by construction as a property of the least-squares technique.
[b] The Hosmer-Lemeshow test statistic is based on median instead of mean predicted costs, because the mean cannot be calculated as a consequence of a negative κ parameter estimate.
[c] Heteroscedasticity is allowed for in this version of the Generalized Gamma Model by defining ln(σ) as a linear combination of the REF adjusters age, gender, insurance eligibility and region of which the weights are to be estimated.

2003). The average residual on the untransformed scale of health care costs for the model variants is also presented.

A variant of the Hosmer and Lemeshow (1995) test determines whether the means of the raw scale residuals across all 10 of the deciles are not significantly different from zero. Assuming a significance level of 5% for these F-tests, there appears to be a systematic pattern of bias in the forecasts for all model variants. In other words, all model specifications are rejected in this case. However, note that the OLS model on the untransformed scale and the Gamma model with a log link are not rejected if a 1% level of significance is assumed (p=0.0146 and p=0.0166, respectively). These latter two model specifications therefore fit the observed data most closely based on this criterion.

Applying Pregibon's Link Test (1981, 1982) for goodness of fit, first the model under consideration is estimated. Then the transformed scale prediction and its squared values are included as the only two variables (in addition to a constant)

in a second regression. Linearity may be assumed if the estimated weight of the squared predictions does not differ significantly from zero. This appears to hold for the OLS model on the untransformed scale under the assumption of a 5% level of significance. Linearity is rejected for all other model variants, even if a 1% level of significance is assumed.

Ramsey's RESET Test (1969) differs from Pregibon's Link Test (1981, 1982) in the sense that the third and fourth powers of the transformed scale predictions are added to the second regression to determine linearity of the residuals at the transformed scale. From Table 7.10 it follows that conclusions to be drawn do not differ from those based on Pregibon's Link Test.

The fourth test uses Pearson correlations between the untransformed scale residuals and the set of explanatory variables. If a tabulated correlation is significantly different from zero, then the corresponding model is providing a biased estimation of REF predicted costs. From Table 7.10 it follows that untransformed scale predictions and residuals are positively correlated under the OLS model for the log transformed values of the dependent variable. There appears to be a negative relationship for the model specifications estimated under the GLM framework. Therefore, all model estimations except the OLS model of untransformed costs generate biased estimations of REF predicted costs. Note that the test result equals zero by construction for the OLS model of untransformed health care costs.

The conclusion must be that the OLS model for untransformed observed costs is the preferred model choice. As an alternative, the standard gamma model specification can only be judged applicable at a 1% level of significance based on the outcome of the Hosmer-Lemeshow test. In any case, the OLS model for log transformed health care costs is among the least preferred alternatives. Also, for the generalized Gamma model, the ancillary parameters are tested such that a particular distribution can possibly be identified. The results of these tests do not alter the conclusions drawn before, the model specifications are rejected in all cases. Note, however, that the $\chi 2$ statistic value of this test for identifying the appropriate distribution is lowest for the log-normal distribution and therefore this distribution most closely fits that of the observed data out of the four distributions tested here.[140]

Summarizing the test results derived above: (1) the log link is tested positively based on the Box-Cox test (Table 7.8); (2) the gamma distribution is tested

140. The Appendix A7.4 shows the results of a sensitivity analysis (Tables A7.4, A7.5 and A7.6). It appears that the Gamma distribution and the log link model specification is not more probable if outliers are excluded from the study sample (i.e. observations for enrollees with observed costs above 50,000 euro in 2002).

Table 7.11: REF predicted costs compared to normative costs 2002 for unadjusted REF weights and REF weights derived in a GLM framework, for survey respondents grouped by the S-type adjusters from the normative equation [a]

		REF predicted costs – normative costs (pipy)		REF predicted costs / normative costs	
Subgroups	**Size of subgroup**	**REF weights**	**REF weights estimated by GLM**	**REF weights**	**REF weights estimated by GLM**
Q1 PF scores	25.0%	-752 *	-521 *	0.799 *	0.861 *
Q1 RP scores	25.0%	-621 *	-403 *	0.823 *	0.885 *
Q1 BP scores	25.0%	-628 *	-468 *	0.800 *	0.851 *
Q1 GH scores	25.0%	-599 *	-367 *	0.829 *	0.895 *
Q1 VT scores	25.0%	-419 *	-232 *	0.853 *	0.919 *
Q1 SF scores	25.0%	-514 *	-307 *	0.826 *	0.896 *
Q1 RE scores	25.0%	-318 *	-217 *	0.882 *	0.919 *
Q1 MH scores	25.0%	-221 *	-170 *	0.905 *	0.927 *
Q2 PF scores	25.0%	124 *	47	1.073 *	1.028
Q2 RP scores	25.0%	77 *	2	1.050 *	1.001
Q2 BP scores	25.0%	40	28	1.024	1.017
Q2 GH scores	25.0%	46 **	-35	1.028 **	0.979
Q2 VT scores	25.0%	-3	-46	0.999	0.974
Q2 SF scores	25.0%	1	-61	1.000	0.967
Q2 RE scores	25.0%	103 *	68 *	1.070 *	1.046 *
Q2 MH scores	25.0%	4	52	1.002	1.028
Q3 PF scores	25.0%	292 *	197 *	1.308 *	1.208 *
Q3 RP scores	25.0%	288 *	208 *	1.272 *	1.197 *
Q3 BP scores	25.0%	252 *	177 *	1.221 *	1.155 *
Q3 GH scores	25.0%	225 *	133 *	1.200 *	1.118 *
Q3 VT scores	25.0%	163 *	105 *	1.126 *	1.081 *
Q3 SF scores	25.0%	250 *	180 *	1.211 *	1.151 *
Q3 RE scores	25.0%	118 *	106 *	1.074 *	1.067 *
Q3 MH scores	25.0%	93 *	62 **	1.065 *	1.043 **
Q4 PF scores	25.0%	336 *	277 *	1.541 *	1.446 *
Q4 RP scores	25.0%	256 *	192 *	1.278 *	1.209 *
Q4 BP scores	25.0%	336 *	263 *	1.319 *	1.249 *
Q4 GH scores	25.0%	328 *	269 *	1.457 *	1.375 *
Q4 VT scores	25.0%	259 *	173 *	1.232 *	1.155 *
Q4 SF scores	25.0%	263 *	188 *	1.254 *	1.182 *
Q4 RE scores	25.0%	97 *	43	1.076 *	1.034
Q4 MH scores	25.0%	123 *	56 **	1.087 *	1.039 **
Number of self-reported OECD limitations [b]					
0	77.8%	204 *	136 *	1.168 *	1.112 *

1	10.3%	-370 *	-336 *	0.873 *	0.885 *
2	4.9%	-916 *	-635 *	0.768 *	0.839 *
3+	5.2%	-1445 *	-785 *	0.726 *	0.851 *
Imputed	1.8%	-13	52	0.994	1.026
Number of self-reported chronic conditions [b]					
0	90.4%	93 *	37 *	1.066 *	1.026 *
1	8.1%	-778 *	-543 *	0.831 *	0.882 *
2	1.3%	-1268 *	809 *	0.806 *	1.124 *
3+	0.2%	-2511 *	177	0.726 *	1.019
Total	100.0%	0	0	1.000	1.000

* Difference between average predicted and normative costs is statistically significant (two-sided t-test, p <= 0.05).

[a] The weighted average of the absolute values of the differences between REF predicted costs and normative costs for the tabulated subgroups equals 171. If the subgroups defined by the S-type adjusters age, gender, PCGs and DCGs are also taken into account (see the Appendix A7.3) then this figure equals 172.

[b] The sizes of these subgroups sum up to 100%.

positively by the modified Park test (Table 7.9 and Figure 7.2); (3) the OLS model for the untransformed health care costs is tested as the most appropriate model if the generalized gamma modeling strategy is applied (Table 7.10). The first two test results favor a log link specification under the assumption of a gamma distribution, however the latter test result favors the standard approach to estimate the 2004 Dutch REF equation (2.1) applied in practice. As the standard approach is already studied in Chapter Six, REF predicted costs are derived in a GLM framework under the assumption of a log link and a gamma distribution hereafter. In the end, given the conflicting test results, the ultimate test is to determine whether this alternative specification generates REF predicted costs that better resemble normative costs than the standard OLS approach.

The gap between REF predicted costs and normative costs is substantially reduced if the REF weights are estimated in a GLM framework for all subgroups defined by the S-type adjusters presented in Table 7.11, if only statistically significant effects are taken into account ($p<0.05$).[141] On the other hand, an increase of this gap can be observed for almost all subgroups defined by age, gender, PCGs

141. In order to be able to assume a gamma distribution, all expenses have to be non-zero. Therefore, one Euro is added to the observed costs of all insured people. This implies a shift of the distribution that does not alter its higher than first order moments. After the estimations have been performed, one Euro is subtracted both from observed costs and normative costs resulting from equation (2.4). Furthermore, as the mean of raw scale REF predicted costs is about 7% higher than the mean of observed costs, REF predicted costs are calibrated by this 7% afterwards.

and DCGs.[142] In particular, REF predicted costs under a GLM framework appear to be above normative costs for the elderly and for those who belong to any of the PCGs and DCGs. This pattern is also observed for the subgroups of insured with two or more self-reported chronic conditions. Note that this phenomenon did not occur when adding PMCGs, MDCGs and MPCGs to the REF equation (see Section 7.1) or after ex-post risk-sharing is applied (see Section 7.2).

If all subgroups defined by the S-type adjusters are taken into account, including age, gender, PCGs and DCGs, then the weighted average of the absolute values of the reported deviations from normative costs following equation (2.6) equals 172. Therefore, under the GLM approach, the policy goals of the Dutch government can be achieved up to an extent of (1-172/687) x 100% = 75.0%.[143] Note that this performance figure is smaller than that after adding the risk adjusters to the REF equation (see Section 7.1), as well as after ex-post risk-sharing is applied (see Section 7.2).[144] However, given this improvement of the performance figure compared to the 71.2% that holds with respect to the 2004 Dutch REF equation, the estimation of the REF weights in a GLM framework in order to capture the multiplicative effects between the REF adjusters (possibly in combination with PMCGs, MDCGs and MPCGs added and/or an ex-post risk sharing scheme) deserves attention in future research.[145]

7.4 Conclusions

In Chapter Six, a normative test for the effectiveness of the risk-adjusted premium subsidies was applied to the 2004 Dutch REF model, given the normative

142. See the Appendix A7.3 for the results with respect to age, gender, the PCGs and DCGs.

143. The figure of 687 can be found in Table A6.1, footnote a.

144. If the subgroups defined by the S-type adjusters age, gender, PCGs and DCGs are not taken into account, then the weighted average of the absolute values of the reported deviations from normative costs equals 251 for the 2004 Dutch REF equation, 199 if PMCGs, MDCGs and MPCGs are added as new REF adjusters, 180 after ex-post risk sharing is applied, and 171 under the GLM approach. These figures are reported in the notes to Tables 6.6, 7.6, 7.7 and 7.11, respectively. Therefore, if the subgroups defined by the S-type adjusters age, gender, PCGs and DCGs are not taken into account, then the performance appears to be even better under the GLM approach than when ex-post risk-sharing is applied.

145. If the deviations from normative costs for the subgroups defined by age, gender, the PCGs and DCGs are ignored, then the summary statistics about the reductions of these deviations are 20.0%, 29.9% and 32.7% for the models proposed in Section 7.1, 7.2 and 7.3, respectively. Note that in this case the reduction appears to be largest if the REF weights are estimated in a GLM framework.

costs as derived in Chapter Five. From the results, it appeared that there remains ample room to improve upon REF predicted costs in order to ultimately fully satisfy the criterion of effectiveness. By construction, a normative adjustment of REF regression weights did completely remove the gap between REF predicted costs and normative costs for the subgroups defined by the REF adjusters. However, of course, for the subgroups defined by the S-type adjusters the gap remained. Therefore, the extent to which other specifications of the REF model may improve the effectiveness of the risk-adjusted premium subsidies is determined in this chapter.

In section 7.1, so-called PMCGs, MDCGs and MPCGs were normatively tested for their contribution as additional REF adjusters to diminish the gap between REF predicted costs and normative costs. Paramedic diagnostic referral codes of chronic diseases are used as indicators of physical limitations (PMCGs); types of medical devices as indicators of functional problems (MDCGs); and pharmaceutical drugs acting on the nervous system as indicators for mental diseases (MPCGs). These potential new REF adjusters are derived from the 2001 Agis claims data. If added to the REF adjusters that are already included in the 2004 Dutch REF scheme, it shows that they substantially reduce the gap between REF predicted costs and normative costs. Nevertheless, part of the gap with normative costs remains for most subgroups defined by the S-type adjusters.

In Section 7.2, risk sharing was tested as a supplement to incomplete and/or imperfect REF adjusters. An analogue of the 2004 Dutch risk sharing scheme have been applied, which essentially boils down to a 90% retrospective reimbursement of actual health care costs above a threshold of € 12,500. Ex-post risk sharing turns out to close the gap between REF predicted costs and normative costs to a large extent, even better than the addition of the PMCGs, MDCGs and MPCGs to the REF equation. Therefore, ex-post risk sharing probably remains an important supplement to incomplete and/or imperfect REF adjusters, even after the PMCGs, MDCGs and MPCGs would be added to the 2004 Dutch REF equation.

In Section 7.3, REF weights are derived in a GLM framework under the assumption of a Gamma error distribution and a log link between REF predicted costs and the REF adjusters. The gap between REF predicted costs and normative costs is substantially reduced if the REF weights are estimated in a GLM framework for almost all subgroups defined by the S-type adjusters, albeit a smaller reduction than under the variants described in Sections 7.1 and 7.2. Nonetheless, the contribution to the effectiveness of the risk-adjusted premium subsidies is still significant, therefore additional research on multiplicative effects between expenses and REF adjusters that are not captured in the conventional REF formula

(possibly in combination with PMCGs, MDCGs, MPCGs and/or ex-post risk sharing schemes) is recommended.

In summary, for the subgroups defined by the S-type adjusters it turns out that REF predicted costs most closely resemble normative costs if a specific type of retrospective risk sharing is applied. Furthermore, the use of PMCGs, MDCGs and PMCGs as well as the application of alternative functional and distributional assumptions about the error terms deserve attention in future research. Note that the exercises in this chapter are merely illustrations of how to apply the normative approach in practice; other model specifications can be studied with this approach as well.

APPENDIX A7.1 A SUPPLEMENT TO SECTION 7.1

Table A7.1: REF predicted costs compared to normative costs 2002 given the old and new REF weights (see Table 7.5), for survey respondents grouped by the S-type adjusters from the normative equation: age, gender and the PCGs and DCGs which are not rank-ordered

		REF predicted costs – normative costs (pipy)		REF predicted costs / normative costs	
Subgroups	**Size of subgroup**	**Old REF weights**	**New REF weights**	**Old REF weights**	**New REF weights**
M 15-24	4.2%	0	0	1.000	1.000
M 25-34	7.1%	0	0	1.000	1.000
M 35-44	6.9%	0	0	1.000	1.000
M 45-54	6.3%	0	0	1.000	1.000
M 55-64	5.7%	0	0	1.000	1.000
M 65-74	5.0%	0	0	1.000	1.000
M 75-84	2.3%	0	0	1.000	1.000
M >=85	0.3%	0	0	1.000	1.000
F 15-24	6.2%	0	0	1.000	1.000
F 25-34	11.3%	0	0	1.000	1.000
F 35-44	12.8%	0	0	1.000	1.000
F 45-54	11.3%	0	0	1.000	1.000
F 55-64	8.9%	0	0	1.000	1.000
F 65-74	7.2%	0	0	1.000	1.000
F 75-84	4.0%	0	0	1.000	1.000
F >=85	0.7%	0	0	1.000	1.000
No PCG	91.2%	9	9	1.006	1.006
PCG01	4.0%	-422 *	-333 *	0.917 *	0.934 *
PCG02	0.5%	-65	-30	0.986	0.993
PCG03	0.2%	482	498	1.130	1.134
PCG04	3.1%	-261	-251	0.960	0.962
PCG05	0.3%	-12	-8	0.998	0.999
PCG06	0.1%	-40	2	0.994	1.000
PCG07	1.2%	-48	-44	0.992	0.993
PCG08	0.1%	-208	-212	0.983	0.982
PCG09	0.0%	0	0	1.000	1.000
PCG10	0.1%	0	0	1.000	1.000
PCG11	0.1%	0	0	1.000	1.000
PCG12	0.0%	0	0	1.000	1.000
No DCG	97.2%	0	0	1.000	1.000
DCG01	0.5%	-185	-186	0.963	0.962
DCG02	0.6%	-147	-183	0.984	0.980
DCG03	0.5%	-1482 *	-1407 *	0.829 *	0.838 *
DCG04	0.5%	-1109	-877	0.900	0.921

Subgroups	Size of subgroup	REF predicted costs – normative costs (pipy)		REF predicted costs / normative costs	
		Old REF weights	New REF weights	Old REF weights	New REF weights
DCG05	0.3%	261	309	1.031	1.037
DCG06	0.1%	-98	32	0.991	1.003
DCG07	0.3%	78	127	1.007	1.012
DCG08	0.2%	123	165	1.009	1.012
DCG09	0.0%	0	0	1.000	1.000
DCG10	0.1%	-290	-259	0.988	0.989
DCG11	0.1%	0	0	1.000	1.000
DCG12	0.1%	-2717	-2639	0.886	0.890
DCG13	0.0%	0	0	1.000	1.000
Total	100.0%	0	0	1.000	1.000

* Difference between average predicted and normative costs is statistically significant (two-sided t-test, p <= 0.05).

APPENDIX A7.2 A SUPPLEMENT TO SECTION 7.2

Table A7.2: REF predicted costs with and without risk sharing compared to normative costs 2002, for survey respondents grouped by the S-type adjusters from the normative equation: age, gender and the PCGs and DCGs which are not rank-ordered

Subgroups	Size of subgroup	REF predicted costs – normative costs (pipy)		REF predicted costs / normative costs	
		Without risk sharing	Including risk sharing	Without risk sharing	Including risk sharing
M 15-24	4.2%	0	0	1.000	1.000
M 25-34	7.1%	0	0	1.000	1.000
M 35-44	6.9%	0	0	1.000	1.000
M 45-54	6.3%	0	0	1.000	1.000
M 55-64	5.7%	0	0	1.000	1.000
M 65-74	5.0%	0	0	1.000	1.000
M 75-84	2.3%	0	0	1.000	1.000
M >=85	0.3%	0	0	1.000	1.000
F 15-24	6.2%	0	0	1.000	1.000
F 25-34	11.3%	0	0	1.000	1.000
F 35-44	12.8%	0	0	1.000	1.000
F 45-54	11.3%	0	0	1.000	1.000
F 55-64	8.9%	0	0	1.000	1.000
F 65-74	7.2%	0	0	1.000	1.000
F 75-84	4.0%	0	0	1.000	1.000
F >=85	0.7%	0	0	1.000	1.000
No PCG	91.2%	9	9	1.006	1.006

		REF predicted costs – normative costs (pipy)		REF predicted costs / normative costs	
Subgroups	**Size of subgroup**	**Without risk sharing**	**Including risk sharing**	**Without risk sharing**	**Including risk sharing**
PCG01	4.0%	-422 *	-229 *	0.917 *	0.955 *
PCG02	0.5%	-65	-99	0.986	0.978
PCG03	0.2%	482	18	1.130	1.005
PCG04	3.1%	-261	-95	0.960	0.986
PCG05	0.3%	-12	-31	0.998	0.995
PCG06	0.1%	-40	-79	0.994	0.988
PCG07	1.2%	-48	-17	0.992	0.997
PCG08	0.1%	-208	-66	0.983	0.995
PCG09	0.0%	0	0	1.000	1.000
PCG10	0.1%	0	0	1.000	1.000
PCG11	0.1%	0	0	1.000	1.000
PCG12	0.0%	0	0	1.000	1.000
No DCG	97.2%	0	0	1.000	1.000
DCG01	0.5%	-185	-5	0.963	0.999
DCG02	0.6%	-147	-159	0.984	0.983
DCG03	0.5%	-1482 *	-221	0.829 *	0.975
DCG04	0.5%	-1109	-176	0.900	0.984
DCG05	0.3%	261	135	1.031	1.016
DCG06	0.1%	-98	-22	0.991	0.998
DCG07	0.3%	78	-165	1.007	0.984
DCG08	0.2%	123	16	1.009	1.001
DCG09	0.0%	0	0	1.000	1.000
DCG10	0.1%	-290	-68	0.988	0.997
DCG11	0.1%	0	0	1.000	1.000
DCG12	0.1%	-2717	-171	0.886	0.993
DCG13	0.0%	0	0	1.000	1.000
Total	100.0%	0	0	1.000	1.000

* Difference between average predicted and normative costs is statistically significant (two-sided t-test, $p <= 0.05$).

APPENDIX A7.3 A SUPPLEMENT TO SECTION 7.3

Table A7.3: REF predicted costs compared to normative costs 2002 for unadjusted REF weights and REF weights derived in a GLM framework, for survey respondents grouped by the S-type adjusters from the normative equation: age, gender and the PCGs and DCGs which are not rank-ordered

		REF predicted costs – normative costs (pipy)		REF predicted costs / normative costs	
Subgroups	**Size of subgroup**	**REF weights**	**REF weights estimated by GLM**	**REF weights**	**REF weights estimated by GLM**
M 15-24	4.2%	0	-46	1.000	0.944
M 25-34	7.1%	0	-4	1.000	0.994
M 35-44	6.9%	0	-109 *	1.000	0.857 *
M 45-54	6.3%	0	-234 *	1.000	0.877 *
M 55-64	5.7%	0	-27	1.000	0.989
M 65-74	5.0%	0	201	1.000	1.053
M 75-84	2.3%	0	765 *	1.000	1.152 *
M >=85	0.3%	0	384 *	1.000	1.103 *
F 15-24	6.2%	0	-57	1.000	0.930
F 25-34	11.3%	0	-77 *	1.000	0.938 *
F 35-44	12.8%	0	-46	1.000	0.955
F 45-54	11.3%	0	-143 *	1.000	0.901 *
F 55-64	8.9%	0	124 *	1.000	1.067 *
F 65-74	7.2%	0	19	1.000	1.007
F 75-84	4.0%	0	379 *	1.000	1.089 *
F >=85	0.7%	0	611 *	1.000	1.168 *
No PCG	91.2%	9	-105 *	1.006	0.924 *
PCG01	4.0%	-422 *	176	0.917 *	1.035
PCG02	0.5%	-65	512 *	0.986	1.114 *
PCG03	0.2%	482	1180	1.130	1.318
PCG04	3.1%	-261	1375 *	0.960	1.209 *
PCG05	0.3%	-12	1152 *	0.998	1.193 *
PCG06	0.1%	-40	349	0.994	1.054
PCG07	1.2%	-48	1557 *	0.992	1.251 *
PCG08	0.1%	-208	2790 *	0.983	1.234 *
PCG09	0.0%	0	3432	1.000	1.459
PCG10	0.1%	0	8152 *	1.000	1.694 *
PCG11	0.1%	0	4093 *	1.000	1.308 *
PCG12	0.0%	0	39992 *	1.000	1.996 *
No DCG	97.2%	0	-106 *	1.000	0.931 *
DCG01	0.5%	-185	944 *	0.963	1.190 *
DCG02	0.6%	-147	6923 *	0.984	1.755 *
DCG03	0.5%	-1482 *	619	0.829 *	1.071

DCG04	0.5%	-1109	657	0.900	1.059
DCG05	0.3%	261	-184	1.031	0.978
DCG06	0.1%	-98	1393 *	0.991	1.125 *
DCG07	0.3%	78	3104 *	1.007	1.294 *
DCG08	0.2%	123	3612 *	1.009	1.269 *
DCG09	0.0%	0	9097 *	1.000	1.823 *
DCG10	0.1%	-290	5326 *	0.988	1.225 *
DCG11	0.1%	0	8542 *	1.000	1.566 *
DCG12	0.1%	-2717	12098 *	0.886	1.506 *
DCG13	0.0%	0	46473 *	1.000	1.552 *
Total	100.0%	0	0	1.000	1.000

* Difference between average predicted and normative costs is statistically significant (two-sided t-test, p <= 0.05).

Appendix A7.4 Normative test results of Section 7.3 after the exclusion of outliers

The Tables A7.4, A7.5 and A7.6 in this appendix are comparable to Tables 7.8, 7.9 and 7.10, respectively. The only difference is that the results below are based on a study sample after the exclusion of 155 outliers from the study sample of 18,617 observations. The 155 outliers are observations for respondents with total health care costs above 50,000 euro in 2002.

From Table A7.4 it follows that the tested link functions do not suffice for modeling purposes, which matches the conclusion drawn from Table 7.8. Note that in case of a log link and a linear link the $\chi 2(54)$ value is substantially reduced after the exclusion of 155 outliers.

Table A7.4: Box-Cox test of null hypothesis $H_0 : \lambda = \hat{\lambda}$, after the exclusion of 155 outliers from the study sample of 18,617 observations for respondents with total health care costs above 50,000 euro in 2002 [a]

$\hat{\lambda}$	**Type of link function**	**Likelihood ratio test $\chi 2(54)$**
-1	Reciprocal link	1.2e+05 (p<0.001)
0	Log link	7.73 (p<0.005)
1	Linear link	39493.31 (p<0.001)

[a] The Box-Cox test is performed without applying sampling weights to the individual observations.

In Table A7.5 the results of the modified Park test are presented. Under the assumption of a log link, only the gamma distribution seems to be able to describe the

relationship between squared residuals and REF predicted costs. This is in line with the conclusion drawn from Table 7.9, except that in this case it also holds under the assumption of a 5% level of significance if equation (7.11) is estimated by OLS.

Table A7.5: Modified Park test to determine choice of preferred type of distribution in a GLM framework, after the exclusion of 155 outliers from the study sample of 18,617 observations for respondents with total health care costs above 50,000 euro in 2002.

Hypothesis about estimated coefficient γ	Type of model distribution	F(1, 18615), given GLM estimates [a]	F(1, 18615), given OLS estimates [b]
0	Gaussian	7130.79 (p<0.0001)	6392.69 (p<0.0001)
1	Poisson	1717.76 (p<0.0001)	1570.10 (p<0.0001)
2	Gamma	2.41 (p=0.1206)	0.50 (p=0.4806)
3	Wald or inverse Gaussian	1984.73 (p<0.0001)	1683.89 (p<0.0001)

[a] The $\hat{\varepsilon}_t^2$ and $\hat{\alpha}_j$, j=0,1,…,J parameters in the modified Park test equation (7.10) result from GLM estimation of equation (7.6), under the assumption of a log link and a gamma distribution.
[b] The $\hat{\varepsilon}_t^2$ and $\hat{\alpha}_j$, j=0,1,…,J parameters in the modified Park test equation (7.10) result from OLS estimation of equation (7.6), under the assumption of a log link.

Table A7.6 shows a systematic pattern of bias in the forecasts for all model variants, except for the OLS model of untransformed health care costs at a 5% level of significance (p=0.7492). Note that in Table 7.10, the Gamma model with a log link was also not rejected at a 1% level of significance.

Table A7.6: Goodness of fit on the untransformed scale of 2002 health care costs (+ 1 euro), after the exclusion of 155 outliers from the study sample out of 18,617 observations for respondents with total health care costs above 50,000 euro in 2002.

Estimator	Hosmer-Lemeshow test	Pregibon Link test	Ramsey RESET test	Pearson correlation test	Average residual
OLS for y	0.67 (p=0.7492)	0.01 (p=0.9394)	3.72 (p=0.0109)	0.00 (p=1.0000)	0.00
OLS for ln(y)	398.09 (p<0.0001)	20.15 (p<0.0001)	37.40 (p<0.0001)	0.34 (p<0.0001)	1538.43
Gamma regression	2.63 (p<0.0001)	16.91 (p<0.0001)	102.49 (p<0.0001)	-0.23 (p<0.0001)	-57.11
Weibull regression	4.14 (p<0.0001)	16.28 (p=0.0001)	49.10 (p<0.0001)	-0.27 (p<0.0001)	-1.38
GGM [a]	116.78 (p<0.0001)	12.22 (p=0.0005)	41.36 (p<0.0001)	-0.02 (p=0.0198)	826.52
GGM – heteroscedastic errors [a,b]	114.57 (p<0.0001)	12.85 (p=0.0003)	45.58 (p<0.0001)	-0.03 (p<0.0001)	816.88

Tests for identifying distributions		χ^2 statistic	Degrees of freedom	p-value
Standard gamma	$\kappa = \sigma$	2141.17	1	<0.0001
Log-normal	$\kappa = 0$	54.94	1	<0.0001
Weibull	$\kappa = 1$	1085.73	1	<0.0001
Exponential	$\kappa = \sigma = 1$	3920.43	2	<0.0001

[a] The Hosmer-Lemeshow test statistic is based on median instead of mean predicted costs, because the mean cannot be calculated as a consequence of a negative κ parameter estimate.
[b] Heteroscedasticity is allowed for in this version of the generalized gamma model by defining ln(σ) as a linear combination of the REF adjusters age, gender, insurance eligibility and region of which the weights are to be estimated.

The test results from Pregibon's Link Test (1981, 1982) and Ramsey's RESET Test (1969) are in line with those drawn from Table 7.10, i.e. linearity is rejected for all model variants except for the OLS model on the untransformed scale. However, the RESET test does not reject linearity for the OLS model on the untransformed scale only under the assumption of a 1% instead of 5% level of significance in the present case.

The Pearson correlation test results are in line with those drawn from Table 7.10, except for the GGM correlation being not different from zero at the 1% level of significance. The tests for identifying distributions are rejected in all cases. Note that like in Table 7.10 the χ2 statistic value is lowest for the log-normal distribution.

The conclusion is that the Gamma distribution and log link model specification is not more probable if outliers are excluded from the study sample.

Chapter 8

PREMIUM RATE RESTRICTIONS TO IMPROVE AFFORDABILITY

In this study, it is assumed that premium rate restrictions are absent. However, restrictions of the out-of-pocket premium rates (i.e. premium minus subsidy) are implemented in many countries, also in Dutch practice (see Section 1.2). Rate restrictions are implemented to increase affordability by creating implicit risk-related subsidies from low-risk to high-risk individuals, given that the explicit risk-adjusted premium subsidies induced by the 2004 Dutch REF equation are suboptimal, which was already demonstrated in Chapter Six. However, rate restrictions also create predictable profits and losses at the individual level, providing incentives for selection that may threaten quality of care, affordability, and efficiency. Therefore, a tradeoff between achieving the policy goals of the sponsor and the incentives for selection exists. In this chapter, it is assumed that rate regulation takes the form of community rating.

In Section 8.1 the incentives for selection are quantified by the calculation of the difference between REF predicted costs and observed costs for specific subgroups. This demonstrates the traditional approach to tabulate incentives for selection as a consequence of rate restrictions. The aforementioned tradeoff is made explicit within the normative framework developed in this study.

By Dutch law, the rate restrictions also hold for cost variation caused by N-type risk factors, which creates incentives for selection. These incentives for selection can be removed by an abolishment of the rate restrictions in this respect. In Section 8.2, the incentives for selection appear to be reduced somewhat if premiums are risk-rated across the twelve Dutch provinces, which is allowed under Dutch law since 2006.

8.1 The tradeoff between affordability and selection

Premium rate restrictions can be used as a supplement to risk-adjusted premium subsidies based on an incomplete set of REF adjusters in order to create implicit cross-subsidies among the subgroups defined by the S-type risk factors. Under community rating, implicit cross-subsidies are created for any cost variation that remains uncaptured by these REF adjusters. However, community rating also induces (implicit) cross-subsidies for N-type cost variation which is in direct conflict with the policy goals of the sponsor. At the same time, incentives for selection are created which can have adverse effects on quality of care, affordability, and efficiency of production.

The traditional approach to demonstrate the existence of incentives for selection is the tabulation of predictable profits and losses for specific subgroups of insured people. Note that there are no predictable profits or losses for subgroups defined

by the REF adjusters. This is a direct consequence of applying the ordinary least squares estimation technique to the REF equation (2.1). Therefore, under the traditional approach, no attention is being paid to the REF adjusters in the context of incentives for selection. However, following equation (2.9) some insightful observations can be made with respect to these REF adjusters under the approach developed in this study.

The linear regression property is also observed in Table 8.1, where the difference between REF predicted costs and observed costs is zero for all subgroups of insured people defined by the REF adjuster insurance eligibility. However, as Table 6.7 already showed, the 2004 Dutch REF equation overcompensates disabled insured people, whereas those who are (self-)employed and those on social welfare are undercompensated. This is also observed from the subgroup differences between REF predicted costs and normative costs in the penultimate column of Table 8.1.

The values in the last column of Table 8.1 follow from the identity described by equation (2.9) that holds for the difference between REF predicted costs and observed costs. Given the zero difference between REF predicted costs and observed costs for subgroups defined by the REF adjuster insurance eligibility, the difference between normative costs and observed costs is exactly equal but opposite to the difference between REF predicted costs and normative for the tabulated subgroups. This illustrates the fact that overcompensation of the disabled insured people goes hand in hand with overutilization in practice, for example. This overutilization must be interpreted as being caused by N-type cost variation which ideally should be reflected in the premium rates. Under community rating, however, (implicit) cross-subsidies will be created across the subgroups defined by insurance eligibil-

Table 8.1: A comparison between REF predicted costs, normative costs and observed costs 2002, for subgroups defined by eligibility

Subgroups	Size of subgroup	Observed costs (pipy)	REF predicted – observed costs	REF predicted – normative costs	Normative – observed costs
Disabled	9.0%	3204	0	420 *	-420 *
Employed	59.5%	965	0	-24 **	24
Social welfare	4.1%	1689	0	-327 *	327
Unemployed	4.2%	1597	0	-106	106
Retired	20.5%	3573	0	-5	5
Self-Employed	2.8%	839	0	-162 *	162
Total	100.0%	1753	0	0	0

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

ity for cost variation caused by N-type risk factors. This is in conflict with the policy goals of the sponsor.

From Table 6.15 it followed that application of normatively adjusted REF weights to the REF equation would completely remove the compensation for N-type risk factors from the risk-adjusted premium subsidies which are induced by the imperfect REF adjusters. Given the aforementioned identity, in that case the values in the penultimate and last column of Table 8.1 will be equal to zero for all subgroups defined by insurance eligibility. This shows that (implicit) cross-subsidies for N-type cost variation among these subgroups as created by the rate restrictions can be avoided by applying adjusted REF weights to calculate the (explicit) risk-adjusted premium subsidies. The same conclusion holds with respect to subgroups of insured people defined by the REF adjuster region (results not tabulated).

For subgroups of insured people other than those defined by the REF adjusters, the difference between REF predicted costs and observed costs is not equal to zero by construction. Furthermore, these predictable profits and losses may reflect cost variation caused by S-type risk factors as well as N-type risk factors. These effects can be separated by application of the normative framework developed in this study following equation (2.9) such that the tradeoff between effectiveness of the risk-adjusted premium subsidies and selection is made explicit.

In order to construct subgroups other than those defined by the REF adjusters, additional information is usually collected from a health survey or an insurer's administration. In this study, the construction of subgroups of high-risk insured people is based on information about utilization, health status and diseases and disorders derived from the Agis Health Survey 2001 and prior costs over the period 1997-2001 derived from Agis' claims data.[146]

Subgroups defined by prior utilization

Table 8.2 shows predictable losses for subgroups of survey respondents who reported medical utilization in 2001. For example, a predictable loss of 286 euro results for those who used prescribed drugs during a period of 14 days. These predictable losses are created by the regulation of the premium rates (in combination with open enrollment). If REF predicted costs would match normative costs perfectly, then this predictable loss would be reduced by 146 euro and amount to the 139 euro associated with the difference between normative costs and observed costs. In that case, rate restrictions would induce cross-subsidies for cost

146. The results for subgroups based on socio-economic characteristics are presented in the Appendix A8.1.

Table 8.2: A comparison between REF predicted costs, normative costs and observed costs 2002, defined by subgroups of survey respondents who reported medical utilization in 2001

Subgroups	Size of subgroup	Observed costs (pipy)	REF predicted – observed costs	REF predicted – normative costs	Normative – observed costs
Contact with GP	75.9%	1971	-83 **	-68 *	-15
Contact with specialist	39.8%	2806	-387 *	-203 *	-184 *
Hospitalization	7.6%	4393	-447 **	-283 *	-164
Paramedic contact	18.2%	2409	-423 *	-283 *	-140
RIAGG contact	19.0%	1761	-127	-10	-117
Alternative care practitioner	11.3%	1841	-253 *	-171 *	-82
Prescribed drugs	46.9%	2776	-286 *	-146 *	-139 **
Non-prescribed drugs	26.0%	1649	-51	-78 *	28

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

variation caused by N-type risk factors alone. Given the imperfect set of REF adjusters, however, rate restrictions also induce implicit cross-subsidies to correct for the undercompensation from the explicit risk-adjusted premium subsidies for cost variation caused by S-type risk factors. Analogously, the predictable loss of 387 euro for insured people who visited a medical specialist can be decomposed into 203 euro which is cross-subsidized by rate restrictions for S-type risk factors and 184 euro cost variation caused by N-type risk factors. Overall, for the subgroups presented in Table 8.2, rate restrictions create implicit cross-subsidies for cost variation caused by S-type risk factors more than for cost variation caused by N-type risk factors: a weighted average of 59.7% of the predictable losses for the tabulated subgroups can be attributed to undercompensations due to incomplete REF adjusters, where the weights are the corresponding sizes of these subgroups.

Subgroups defined by health status, diseases and conditions

Table 8.3 shows predictable losses for subgroups of survey respondents who reported long-term diseases, psychological distress or functional limitations in 2001 and those for whom obesitas could be determined. These subgroups also include survey respondents who indicated that they are not under treatment or control by a doctor and those who do not have the reported health complaints any more. With only a few exceptions, Table 8.3 shows that predictable losses reflect cost variation caused mainly by S-type risk factors due to incomplete REF adjusters. Therefore, rate restrictions rightly induce implicit cross-subsidies for S-type cost variation. The extent to which rate restrictions induce cross-subsidies

Table 8.3: A comparison between REF predicted costs, normative costs and observed costs 2002, defined by subgroups of survey respondents who reported long-term diseases or psychological distress, functional limitations, and obesitas in 2001, including those survey respondents who did not indicate still having complaints or being under treatment

Subgroups	Size of subgroup	Observed costs (pipy)	REF predicted – observed costs	REF predicted – normative costs	Normative – observed costs
Self-reported long-term diseases					
Diabetes mellitus (Type I and II)	4.3%	4563	-609 *	-734 *	124
Stroke, brain haemorrhage/ infarction (ever)	2.5%	4910	-825 **	-786 *	-39
Myocardial infarction (ever)	3.3%	5457	-956 *	-620 *	-336
Other serious heart disease	2.3%	5197	-425	-866 *	441
Some type of (malignant) cancer (ever)	4.7%	4363	-852 *	-563 *	-288
Migraine or serious headache regularly	22.2%	1619	-93	-205 *	113
Hypertension	14.5%	3264	-489 *	-267 *	-222
Vascular constriction (stomach, legs)	4.0%	4601	-905 *	-547 *	-358
Asthma, COPD	7.9%	3386	-401 *	-376 *	-25
Psoriasis	1.8%	2863	-517	-120	-397
Chronic dermatitis	5.3%	1687	6	-134 *	140
Dizziness when falling down	6.1%	2988	-418 *	-565 *	147
Intestinal obstructions (> 3 months)	4.2%	3261	-796 *	-475 *	-321
Urinary incontinence	7.5%	3503	-844 *	-403 *	-440 *
Serious/persistent back problem	14.6%	2524	-303 *	-466 *	162
Osteoarthritis (hip/knees)	15.5%	3144	-444 *	-447 *	2
Chronic joint inflammation	5.8%	3566	-910 *	-574 *	-336
Other serious/persistent injury (neck, shoulder)	14.8%	2376	-225 *	-355 *	130
Other serious/persistent injury (elbow, wrist, hand)	9.5%	2686	-439 *	-484 *	45
Other prolonged disease/disorder	11.9%	3278	-590 *	-397 *	-193
Self-reported psychological distress					
Fearful or afraid (for 2 weeks)	30.4%	2139	-135	-139 *	4
Downhearted or blue (for 2 weeks)	30.7%	1956	-97	-105 *	8
Either fearful/afraid or downhearted/blue (for 2 weeks)	39.3%	2000	-89	-114 *	24
Self-reported functional limitations					
Stayed in bed (during past 6 months)	37.5%	2011	-205 *	-211 *	6

Subgroups	Size of subgroup	Observed costs (pipy)	REF predicted – observed costs	REF predicted – normative costs	Normative – observed costs
Enrollees with one or more OECD limitations	20.5%	3755	-765 *	-775 *	10
Enrollees with OECD auditive impairment	4.7%	3117	-351	-743 *	392 **
Enrollees with OECD visual impairment	7.0%	3115	-306 **	-692 *	386 *
Enrollees with OECD mobility impairment	14.5%	4409	-1086 *	-994 *	-92
Self-reported lifestyle					
Enrollees with obesitas	11.6%	2244	-159	-298 *	139

Note: Survey respondents assigned to the subgroups diabetes mellitus (Type I and II), myocardial infarction or other serious heart disease, or asthma/COPD do not necessarily coincide with the subgroups of enrollees classified in PCG07=Diabetes Type I, PCG04=Cardiac disease and PCG01=Asthma/COPD, respectively. The former classification is based on self-reports, the latter on administrative information on the use of pharmaceutical drugs.

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

for N-type cost variation is smaller than the intended increase of effectiveness of the risk-adjusted premium subsidies for these subgroups of insured people. The exceptions are the subgroups of insured people with psoriasis, chronic dermatitus and urinary incontinence.[147]

Subgroups defined by prior costs

Table 8.4 shows predictable profits and losses for subgroups defined by the number of years that enrollees belong to the top 25% of total expenses within each year prior to 2002. For the majority of these subgroups of insured people, rate restrictions appear to create implicit cross-subsidies for cost variation caused mainly by N-type risk factors. For example, 498 euro of the predictable loss of 2025 euro for the subgroup of survey respondents who did belong to the top 25% category of total health care costs for five consecutive years must be attributed to S-type risk factors due to an incomplete set of REF adjusters and 1527 euro must be attributed to N-type risk factors. For the subgroups of survey respondents who belonged to the top 25% category of total health care costs for one and three

147. The differences between REF predicted costs and normative costs are mostly statistically significantly different from zero. This is caused by the limited variation at the individual level of REF predicted costs and normative costs relative to observed costs.

Table 8.4: A comparison between REF predicted costs, normative costs and observed costs 2002, defined by the number of years that enrollees belong to the top 25% of total expenses within each year, prior to 2002

Number of years prior to 2002	Size of subgroup	Observed costs (pipy)	REF predicted – observed costs	REF predicted – normative costs	Normative – observed costs
Enrolled < 5 years	13.9%	997	141 *	11	130 *
0	40.6%	742	408 *	153 *	255 *
1	17.8%	1583	26	37	-11
2	9.6%	1920	149	-52	201 *
3	5.8%	2613	-166	-219 *	53
4	4.3%	3748	-767 *	-301 *	-466 **
5	8.0%	6690	-2025 *	-498 *	-1527 *
Total	100.0%	1753	0	0	0

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

years, rate restrictions appear to create implicit cross-subsidies for cost variation mainly caused by S-type risk factors.

8.2 Risk rating across Dutch provinces

Premium rate restrictions induce implicit cross-subsidies for cost variation which is caused by S-type risk factors but not captured by the incomplete set of REF adjusters. However, implementation of rate restrictions comes at the risk of undesirable cross-subsidies for cost variation caused by N-type risk factors at the same time. Incentives for selection with respect to these N-type risk factors are only avoided if the rate restrictions do not hold with respect to these N-type risk factors.

Table 8.5 shows the implicit cross-subsidies which are induced by the rate restrictions in terms of the difference between REF predicted costs and observed costs. Under the approach developed in this study, this cost variation between provinces can be separated into cost variation caused by S-type risk factors but not captured by the REF adjusters (= the difference between REF predicted costs and normative costs) and cost variation caused by N-type risk factors (= the difference between normative costs and observed costs). From a per-province comparison of these separated effects, it appears that the rate restrictions induce implicit cross-subsidies mainly for cost variation caused by N-type risk factors (the provinces Flevoland and Gelderland are the exceptions). This can be seen by comparing the values in the last column to those in the penultimate column. Given

Table 8.5: Predictable profits and losses caused by premium rate restrictions for subgroups defined by the twelve Dutch provinces, given an incomplete set of REF adjusters

Subgroups	Size of subgroup	Observed costs (pipy)	REF predicted – observed costs	REF predicted – normative costs	Normative – observed costs
Drenthe	0.8%	1973	-262	-176	-86
Flevoland	5.4%	2005	-272	35 *	-307
Friesland	5.5%	1569	190	-70 *	259 **
Gelderland	21.8%	1638	-15	-45	29
Groningen	0.5%	1848	214	-125	339
Limburg	0.2%	1661	-437	-13 *	-424
Noord-Brabant	0.5%	2126	-779	-92 *	-687
Noord-Holland	31.8%	2021	-64	26 *	-90
Overijssel	2.2%	1827	-276	-6 **	-271
Utrecht	27.9%	1579	104	25 *	78
Zeeland	0.2%	893	803	2	801
Zuid-Holland	3.3%	1187	279 **	-32 *	312 **
Total	100.0%	1753	0	0	0

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

the tradeoff between the policy goals of the sponsor on the one hand and the incentives for selection on the other hand, it appears to have been a right decision of the Dutch government to allow risk-rating premiums across the twelve Dutch provinces under Dutch law since 2006. In other words, premiums are community rated per province in this case.

Table 8.6 shows that the cost variation caused by S-type risk factors is largely reduced if the normatively adjusted REF weights are applied (see Chapter Six, Section 6.3). This reduction can be seen by a comparison of the difference between the normatively adjusted REF predicted costs and normative costs in Table 8.6 with the difference of REF predicted costs and normative costs in Table 8.5. It appears that the deviance from normative costs is much smaller in Table 8.6, the exceptions being the provinces Overijssel and Zeeland. The remaining cost variation between provinces is caused by N-type risk factors, see the difference between normative costs and observed costs for each subgroup in the last column. Therefore, premiums which are risk-rated across the twelve Dutch provinces mainly reflect cost variation caused by N-type risk factors, even more so if the normatively adjusted REF weights are applied.

Table 8.6: A comparison between REF predicted costs, normative costs and observed costs 2002, defined by subgroups of survey respondents who reported medical utilization in 2001

Subgroups	Size of subgroup	Observed costs (pipy)	Normatively adjusted REF predicted – observed costs	Normatively adjusted REF predicted – normative costs	Normative – observed costs
Drenthe	0.8%	1973	-164	-78	-86
Flevoland	5.4%	2005	-316	-9	-307
Friesland	5.5%	1569	257 **	-2	259 **
Gelderland	21.8%	1638	27	-2	29
Groningen	0.5%	1848	300	-39	339
Limburg	0.2%	1661	-428	-4	-424
Noord-Brabant	0.5%	2126	-748	-62	-687
Noord-Holland	31.8%	2021	-100	-10	-90
Overijssel	2.2%	1827	-225	46	-271
Utrecht	27.9%	1579	93	15	78
Zeeland	0.2%	893	831	30	801
Zuid-Holland	3.3%	1187	314 **	2	312 **
Total	100.0%	1753	0	0	0

*,** Difference between average normative costs and observed costs is statistically significant (two-sided t-test, * p <= 0.05, ** p <= 0.10).

8.3 Conclusions

In Chapter Six, it was demonstrated that the 2004 Dutch REF equation does not fully satisfy the policy goals of the Dutch government. Premium rate restrictions (and open enrollment) hold in Dutch health insurance in order to guarantee affordability by implicit cross-subsidies for cost variation between the subgroups defined by the S-type risk factors. In Section 8.1 it is demonstrated that rate restrictions do not lead to incentives for selection with respect to the subgroups defined by the REF adjusters. Furthermore, no cross-subsidies for N-type cost variation are induced among these subgroups either, as long as the normatively adjusted REF weights are applied.

Rate restrictions may create predictable profits and losses for subgroups of insured people other than those defined by the REF adjusters. For most of the subgroups defined by self-reported prior medical utilization and self-reported health status, diseases and conditions, the predictable losses mainly reflect cost variation caused by S-type risk factors. Therefore, the rate restrictions rightly induce implicit cross-subsidies for S-type cost variation. The extent to which rate restrictions induce cross-subsidies for N-type cost variation is much smaller than the intended increase of affordability for these subgroups of insured people. It

should be noted, however, that the implicit cross-subsidies induced by the rate restrictions come at the expense of incentives for selection.

The predictable profits and losses for subgroups of insured, defined by the number of years that they belong to the top 25% of total expenses within each year prior to 2002, appear to be mainly caused by N-type risk factors. The prevention of incentives for selection can only be avoided by the abolishment of the rate restrictions with respect to these subgroups.

To avoid the incentives for selection with respect to cost variation caused by N-type risk factors, the rate restrictions should be abolished with respect to such risk factors. Under 2006 Dutch law, risk-rating premiums across the twelve Dutch provinces is allowed. From Section 8.2, it appears that premiums that are risk-rated across the twelve Dutch provinces will mainly reflect cost variation caused by N-type risk factors. This is even more so if the normatively adjusted REF weights are applied. From the perspective of the policy goals of the Dutch government, this result justifies the decision to allow risk-rating premiums across subgroups defined by the twelve Dutch provinces. At the same time, incentives for selection are avoided.

Appendix A8.1 Results for socio-economic subgroups

Table A8.1 shows the difference between REF predicted costs and observed costs for subgroups categorized by self-reported after-tax monthly household income.[148] Although no predictable profit or loss is induced by premium rate restrictions for insured people with self-reported after-tax household income of 5000 Dutch guilders a month and above, this appears to be the net result of overcompensation of 112 euro on average induced by a set of imperfect REF adjusters and overutilization of 134 euro.[149] For those insured people with self-reported after-tax household income from 1000 up to 2000 Dutch guilders a month, average undercompensation equals 121 euro because of the use of imperfect REF adjusters and average underutilization equals 157 euro. This means that the rate restrictions not only increase affordability as intended by the sponsor, but also induce cross-subsidies for cost variation caused by N-type risk factors.

Table A8.1: A comparison between REF predicted costs, normative costs and observed costs 2002, defined by subgroups of survey respondents who reported their after-tax monthly household income in 2001 Dutch guilders [a]

Subgroups	Size of subgroup	Observed costs (pipy)	REF predicted – observed costs	REF predicted – normative costs	Normative – observed costs
Not reported	8.5%	1649	80	5	75
Less than 1000	3.4%	1580	-49	-30	-20
From 1000 up to 2000	17.4%	2352	35	-121 *	157
From 2000 up to 3000	28.6%	2033	-49	-49	0
From 3000 up to 4000	19.4%	1486	101	73	27
From 4000 up to 5000	10.5%	1409	-135	74 *	-209
5000 or more	12.1%	1087	-22	112 *	-134 *
Total	100.0%	1753	0	0	0

* Difference between average normative and observed costs is statistically significant (two-sided t-test, p <= 0.05).

[a] 1 Euro = 2.20371 Dutch guilders

148. Note that in order to become a sickness fund member in 2001, individual gross annual income had to be below € 29,813.

149. The overutilization appears to be largely financed by an improper overcompensation induced by the imperfect REF adjusters from the 2004 Dutch REF equation.

Table A8.2 shows that rate restrictions induce implicit cross-subsidies for cost variation caused by S-type and N-type risk factors between subgroups defined by education.[150]

Table A8.2: REF predicted costs compared to observed costs 2002, for subgroups of survey respondents classified by highest level of successfully finished education in 2001

Subgroups	Size of subgroup	Observed costs (pipy)	REF predicted – observed costs	REF predicted – normative costs	Normative – observed costs
Not reported	9.2%	2310	-12	-165 *	153
Elementary (incl. not finished)	17.6%	2323	-44	-104 *	60
Lower secondary (Lbo/ Mavo/Vmbo)	37.9%	1764	20	25	-5
Higher secondary (Havo/Vwo/Mbo)	22.1%	1234	-3	77 *	-79
Tertiary (Hbo)	7.8%	1296	-12	137 *	-150
Tertiary (University)	3.2%	1207	58	98	-40
Otherwise	2.4%	2286	44	-297 **	341
Total	100.0%	1753	0	0	0

Judged by the differences between REF predicted costs and observed costs, rate restrictions only slightly create incentives for risk selection with respect to subgroups of first- and second generation immigrants. However, Table A8.3 shows that for the first-generation immigrants this is the net result of a combination of undercompensation of 233 euro induced by a set of imperfect REF adjusters and underutilization of 275 euro. The undercompensation means that, given their values of the S-type adjusters, more utilization is expected than is actually observed in practice. The underutilization must be attributed to N-type risk factors in general, for example more (communication) problems than others to get access to needed care. This observation does not hold for second-generation immigrants and non-immigrants.

150. The differences between REF predicted costs and normative costs are mostly statistically significantly different from zero. This is caused by the limited variation at the individual level of REF predicted costs and normative costs relative to observed costs.

Table A8.3: A comparison between REF predicted costs, normative costs and observed costs 2002, for survey respondents who are called first- or second generation immigrants in 2001 [a]

Subgroups	Size of subgroup	Observed costs (pipy)	REF predicted – observed costs	REF predicted – normative costs	Normative – observed costs
First generation	12.2%	1673	43	-233 *	275 *
Second generation	5.4%	1563	-58	16	-74
Others	82.4%	1778	-3	33 **	-36
Total	100.00%	1753	0	0	0

* Difference between average normative costs and observed costs is statistically significant (two-sided t-test, p <= 0.05).

[a] A respondent is called an immigrant if at least one of his/her parents is born outside of The Netherlands. Furthermore, the respondent is called a first-generation immigrant if he/she is also born outside of the Netherlands, otherwise he/she is called a second-generation immigrant.

Chapter 9

Conclusions and discussion

In competitive individual health insurance markets, risk-rated premiums are observed to differ across subgroups of insured people, which are defined by rating factors such as: age, gender, family size, geographic area, occupation, length of contract period, the level of deductible, health status at time of enrollment, health habits (smoking, drinking, exercising) and — via differentiated bonuses for multi-year no-claim — to prior costs (Van de Ven et al. 2000). Financial transfers are needed in order to prohibit any problems of financial access to coverage for those at high risk. The first and best solution to increase financial access to coverage for those at high risk is to find a sponsor who organizes a regulatory system of risk-adjusted premium subsidies (Van de Ven et al. 2000). The financial transfers are then channeled via a so-called Risk Equalization Fund (REF). In that case, price competition is not distorted and therefore incentives for efficiency are not reduced. In all countries that apply risk-adjusted premium subsidies in their health insurance market, the sponsor organizes it in the form of risk equalization among insurers.

Although premiums are rated across many subgroups of insured people, a sponsor may not want to subsidize all premium rate variation observed in practice. The total set of risk factors that insurers use to rate their premiums can be divided in two subsets: the subset of risk factors that cause premium rate variation which the sponsor decides to subsidize, the S(ubsidy)-type risk factors; and the subset that causes premium rate variations which the sponsor does not want to subsidize, the N(on-subsidy)-type risk factors (Van de Ven and Ellis 2000, p. 768-769). In most countries, gender, health status and age (to a certain extent) will probably be considered S-type risk factors. Examples of potential N-type risk factors are: a high propensity for medical consumption, living in a region with high prices and/or overcapacity resulting in supply-induced demand, or using providers with an inefficient practice-style (Van de Ven et al. 2000). The sponsor determines the specific categorization of S-type and N-type risk factors. In case the government takes up the role of the sponsor, this categorization is ultimately determined by value judgments in society.

Under current law, Dutch government has decided to cross-subsidize cost variation between subgroups defined by the so-called S-type risk factors age, gender and health status only (MoHWS 2005, p. 23). Measures of age and gender are available in the administrations of all Dutch insurers and therefore can be included into the Dutch REF model relatively easily. However, the empirical possibilities to construct a health-based REF model at the individual level are rather limited. The 2004 Dutch REF equation (2.1) already contains an impressive range of health-based administrative adjusters, including Pharmacy-based Cost Groups (PCGs) and Diagnostic Cost Groups (DCGs). Currently, it is the most elaborate individual

level risk equalization model in the world. However, the question as to what extent even this extensive set of REF adjusters generates risk-adjusted premium subsidies (or cross-subsidies) that satisfy the policy goals of the Dutch government still remains unanswered. The central question of this study is therefore:

"To what extent does the 2004 Dutch risk equalization model induce risk-adjusted premium subsidies that meet the stated policy goals of the Dutch government and (how) can these subsidies be improved?"

In order to find an answer to this central question, three research questions must be formulated:

1. Given the definition of the basic benefits package, how can we calculate the cross-subsidies as intended by the Dutch government? (Chapters Three, Four and Five)
2. To what extent can the intended cross-subsidies be achieved by the health status measures included in the 2004 Dutch REF equation? (Chapter six)
3. To what extent can the intended cross-subsidies be achieved by alternative specifications of the 2004 Dutch REF model or by premium rate regulations? (Chapters Seven and Eight)

The main contribution of this study is the development and empirical application of a theoretical framework to determine the extent to which REF models induce effective cross-subsidies. In this study it is assumed that there is a periodic open enrollment requirement for a specified benefit package and a system of risk-adjusted premium subsidies. However, premium rates in the competitive individual health insurance markets are assumed not to be regulated.

In this study, a procedure is developed to test whether a given set of REF adjusters adequately compensates for cost variation caused by S-type risk factors (Chapter One). An overview is given of the relevant literature on (mainly) administrative measures of health status currently in use or under study. The methodology applied in subsequent chapters is described in more detail as well as in mathematical terms. Furthermore, guidelines are given on the interpretation of results and on the comparisons that are most relevant to find an answer to the central question of this study (Chapter Two).

For the application of the proposed test procedure, a tailor-made health survey was conducted amongst more than 50,000 sickness fund enrolees, such that the health status profile can be described much more broadly than with the REF adjusters alone (Chapter Three). The health status measures in this study are extensively tested for their completeness, reliability and validity (Chapter Four).

Throughout this study it is assumed that the Dutch government desires cross-subsidies for observed cost variation caused by the S-type risk factors age, gender and health status alone. For a limited sample of insured people, an alternative risk equalization model is then developed at the individual level that captures this observed cost variation as accurately as possible by including all measures of the S-type risk factors available from the administrative data sources and the health survey (Chapter Five). The so-called normative costs that follow from this alternative risk equalization model are then compared to REF predicted costs given the set of REF adjusters included in the 2004 Dutch REF equation: age, gender, insurance eligibility, region, PCGs and DCGs (Chapter Six). Following the same test procedure, cross-subsidies from alternative specifications of the REF model are tested for their effectiveness (Chapter Seven). Lastly, it is shown that an improvement of the REF model should be the preferred strategy to increase financial access to coverage rather than implicit cross-subsidies enforced by premium rate restrictions (Chapter Eight).

9.1 Conclusions

An answer to the first research question

Given the definition of the basic benefits package, the first research question determines how to calculate the cross-subsidies as intended by the Dutch government. In Chapter Five the normative costs are derived for a limited sample of insured people (N=18,617) under the assumption that society desires cross-subsidies for cost differences caused by the S-type risk factors age, gender and health status. The normative costs follow from a linear regression of observed costs on a broad array of health status variables from the health survey and administrative sources under the assumption that these form an adequate reflection of the S-type risk factors.

From the literature on risk adjusters as described in Chapter Two, it appears that the most promising candidates as measures of health status are self-reported measures of perceived health status, functional health status and chronic conditions. In order to guide the specific selection of health status measures, the conceptual model of Ruwaard and Kramers (1997) is applied. The selected health status measures are the eight SF-36 scales physical functioning (PF), role-physical (RP), bodily pain (BP), general health (GH), vitality (VT), social functioning (SF), role-emotional (RE) and mental health (MH). Three categories are based on the number of OECD (auditive, visual and mobility) limitations and three categories are based on the number of specific self-reported chronic conditions. The SF-36 is

a 36-item instrument for measuring health status and outcomes from the patient's point of view; it was designed for use in clinical practice and research, health policy evaluations and general population surveys (Ware and Hays 1988, Aaronson et al. 1998). PCGs and DCGs are finally added to the normative equation, as costs of medical care are not necessarily larger for lower scores on the above mentioned health status indicators (Newhouse 1989).

In Chapter Three the data that is used in this study is described. A broad array of health status measures is obtained by means of a tailor-made health survey which was conducted in 2001 amongst 50,022 Agis members. Gross response to the so-called "Agis Health Survey 2001" was 23,163 (46.3%). For the purpose of this study, 18,617 eligible records are included in the study sample, because valid SF-36 scale scores could be derived and the administrative records from the years 2001 and 2002 appeared to be both valid and available for these respondents. Reliability and validity of the eight SF-36 scales are tested positively in Chapter Four. Furthermore, a panel dataset is derived from the Agis social health insurance administration 1997-2002, as well as additional 2001 data which is kindly made available by Dutch research institutes (APE Public Economics and Prismant, located in The Hague and Utrecht, respectively).

Normative costs are then derived by ordinary least-squares, following equation (2.4) as the average observed costs 2002 for the subgroups defined by the so-called S-type adjusters age and gender, the non-rankordered versions of the PCGs and DCGs, the eight SF-36 scales, the number of OECD limitations, and the number of self-reported chronic conditions. Note that in contrast to the rank-ordered versions of the (rank-ordered) PCGs and DCGs which are included in the 2004 REF equation, in the normative equation, enrollees may be associated with multiple PCGs and DCGs. An answer to the first research question is that cross-subsidies which are in line with the policy goals of the sponsor can be calculated for a limited sample of survey respondents by following the approach developed in this study.

An answer to the second research question

In order to answer the second research question, the extent must be determined to which the health status measures included in the 2004 Dutch REF equation induce the cross-subsidies as intended by the Dutch government. Ideally, cross-subsidies are based on normative costs in order to be exactly in accordance with the policy goals of the sponsor. In that hypothetical case, for the subgroups defined by the S-type adjusters, REF predicted costs are equal to average observed costs. Therefore, the cross-subsidies which are implemented in practice can be tested by comparing REF predicted costs to normative costs for the subgroups defined by the S-type adjusters.

In Chapter Six, a test of the adjusters included in the 2004 Dutch REF equation is performed, given the normative costs as derived in Chapter Five. Following equation (2.6), the weighted average of the deviations of REF predicted costs from normative costs for the subgroups defined by the S-type adjusters equals 198. This number would be 687 in the absence of any risk adjusters in equation (2.1); that is, if REF predicted costs are equal to average observed costs for all insured people. Therefore, 1 – (198/687) x 100% = 71.2% of the cross-subsidies that the Dutch government desires can be achieved by the REF adjusters included in the 2004 Dutch REF equation. This finding answers the second research question.

REF predicted costs and normative costs can also be compared for subgroups defined by the REF adjusters (instead of the S-type adjusters). Deviations of REF predicted costs from normative costs may be attributed to N-type risk factors. REF predicted costs for disabled enrollees appear to be overcompensated for by 420 euros (15.1%), while enrollees on social welfare and self-employed enrollees are under compensated for by 327 euros (16.2%) and 162 euros (16.2%) relative to normative costs, respectively. Insured people living in the first regional cluster of zip codes are overcompensated for by 245 euros (13.6%) relative to normative costs, whereas enrollees living in the regional clusters 6, 7 and 8 are under compensated for by 190 euros (10.7%), 94 euros (5.4%) and 171 euros (11.4%) on average. Therefore, the assumption in the 2004 Dutch REF model that these REF adjusters are "unbiased" measures of health status differences must be rejected; i.e., in practice these REF adjusters lead to compensation for N-type cost variation. In this context it is important to notice that the study sample contains insurance members of only one Dutch health insurer, therefore the research results cannot be treated as representative of Dutch provinces.

Furthermore, it appears that REF predicted costs for self-employed enrollees are lower than those for employed enrollees, although normative costs are not for these subgroups. Apparently, the REF adjuster insurance eligibility fails to adequately capture S-type cost variation for self-employed enrollees. It should be noted that this problem can not be tackled by combining the subgroups of employed and self-employed enrollees into one subgroup, a solution which was heavily debated in the context of the 2004 specification of the Dutch REF equation.[151] In the end, the Dutch government decided to treat employed and self-employed insured people as separate subgroups in the 2004 REF equation, under the assumption that the difference could be attributed mainly to the S-type risk

151. Indeed, this solution would come at the expense of reduced risk-adjusted premium subsidies for employed enrollees.

factor health status. However, this decision has lead to lower premium subsidies for self-employed enrollees which from the results presented in this study cannot be justified given the policy goals as stated by the Dutch government.

Ideally, the gap between REF predicted costs and normative costs is removed such that the REF adjusters no longer induce cross-subsidies for cost variation caused by N-type risk factors. This may be achieved by an adjustment of the REF weights, for example by application of the so-called omitted variables approach or the so-called normative adjustment approach developed in this study. The omitted variables approach to remove N-type bias from the REF weights is proposed by Schokkaert, Dhaene and Van de Voorde (1998) and Schokkaert and Van de Voorde (2004). However, it turns out that following this approach, the reduction of the gap between REF predicted costs and normative costs is rather limited, at least given the study sample and the specific implementation of N-type adjusters in this study. An alternative procedure to adjust the REF weights is to replace them by so-called normatively adjusted REF weights that follow from equation (2.8) such that the bias is completely removed from the unadjusted REF weights.

The adjustment of the REF weights following the omitted variables approach leads to risk-adjusted premium subsidies to be in line with the policy goals of the Dutch government up to an extent of (1-201/687) x 100% = 70.7%. Alternatively, in case of a normative adjustment of the REF weights this performance outcome equals (1-209/687) x 100% = 69.6%. Apparently, a removal of the N-type bias from the REF weights at the same time slightly reduces the amount of S-type cost variation that is captured by the REF adjusters. Therefore, it is recommended to use these adjusted REF weights in a REF equation rather than the original REF weights in risk equalization models if this tradeoff is not too severe.

An answer to the third research question

In order to answer the third research question, the extent must be determined to which effective cross-subsidies can be achieved by alternative specifications of the 2004 Dutch REF model or by premium rate regulation. In Chapter Six, given the set of REF adjusters included in the 2004 Dutch REF equation, an adjustment of the REF weights is applied in order to remove the gap between REF predicted costs and normative costs *for the subgroups defined by the REF adjusters*. However, this procedure does not change the gap *for the subgroups defined by the S-type adjusters* and therefore does not improve the cross-subsidies to better compensate for cost variation caused by S-type risk factors. In Chapter Seven alternative specifications of the 2004 Dutch REF model are tested in order to reduce this gap for the subgroups defined by the S-type adjusters. This procedure may lead to cross-subsidies that are better in line with the policy goals of the sponsor.

As a first example, so-called paramedic cost groups (PMCGs), medical device cost groups (MDCGs) and mental pharmacy-based cost groups (MPCGs) are constructed from administrative data. Paramedic indicators of chronic diseases are used as indicators of physical limitations, medical devices are used as indicators of functional problems and pharmaceutical drugs for treatment of the nervous system as indicators for mental diseases. These potential new REF adjusters are derived from the claims data in the Agis sickness fund administration. If added to the REF adjusters that are already included in the 2004 Dutch REF model, it shows that they substantially reduce the gap between REF predicted costs and normative costs. Following equation (2.6), the weighted average of the absolute deviations of REF predicted costs from normative costs for subgroups of S-type adjusters equals 157. Given that this number would be 687 in the absence of any system of risk-adjusted premium subsidies (1 – 157/687) x 100% = 77.1% of effective cross-subsidies can be achieved after adding the aforementioned REF adjusters to the 2004 Dutch REF equation. Remember that this figure equals 71.2% if these new REF adjusters are not added. Thus, adding PMCGs, MDCGs and MHCGs will lead to a substantial improvement of the risk-adjusted premium subsidies; however, some room for improvement still remains.

As a second example, a specific ex-post risk sharing scheme is tested as a supplement to incomplete and/or imperfect REF adjusters. An analogue of the 2004 Dutch risk sharing scheme is applied: a 100% retrospective reimbursement of production-independent observed hospital costs and a 90% retrospective reimbursement of observed production-dependent hospital costs, medical specialty costs and costs of other health care services above a threshold of € 12,500. In this study, this turns out to close the gap between REF predicted costs and normative costs to a large extent. The weighted average of the deviations of REF predicted costs from normative costs for subgroups of S-type adjusters equals 141 in this case. Therefore, (1 – 141/687) x 100% = 79.5% of the intended cross-subsidies can be achieved as by the aforementioned ex-post risk-sharing arrangement as a supplement to the 2004 Dutch REF equation. Ex-post risk sharing will also remain an important supplement to incomplete and/or imperfect adjusters, even after the PMCGs, MDCGs and MPCGs are added to the REF equation. It should be noted that implementation of an ex-post arrangement in practice reduces the incentives for efficiency and therefore introduces a tradeoff between the improvement of the cross-subsidies and efficiency.

As a third example, REF weights are derived in a GLM framework under the assumption of a Gamma error distribution and a log link between REF predicted costs and the REF adjusters. The discrepancy between REF predicted costs and normative costs is substantially removed for almost all subgroups defined by the

S-type adjusters; albeit this reduction is smaller than under the aforementioned variants of the 2004 Dutch REF equation. The weighted average of the deviations of REF predicted costs from normative costs for subgroups of S-type adjusters equals 172. Therefore, (1 – 172/687) x 100% = 75.0% of effective cross-subsidies can be achieved by changing the statistical specification of the REF model.

Throughout this study, it is assumed that there is a periodic open enrollment requirement for a specified benefit package. Insurers are free to rate their premiums, and a system of risk-adjusted premium subsidies is organized by a sponsor to safeguard financial access to coverage for high-risk insured. Given that none of the specifications of the REF model induces cross-subsidies that fully capture cost variation caused by S-type risk factors, the sponsor may be tempted to impose premium rate regulations in order to safeguard financial access to coverage for high-risk insured anyhow.

Premiums are direct payments which insured people pay to their insurers and their rates can be regulated by the sponsor. Premium rate regulation can take several forms: community-rating per insurer, a ban on certain rating factors or rate-banding (by class). Community-rating is the most extreme variant of rate regulation, because all insured people enrolled with the same insurer are obliged to pay the same premium for a specified benefit package irrespective of their individual risk profiles. However, although premium rate restrictions are intended to create implicit cross-subsidies for cost variation caused by S-type risk factors alone, these may also induce cross-subsidies for cost variation caused by N-type risk factors, which is by definition in conflict with the policy goals of the sponsor. Given the theoretical framework developed in this study, it is possible to determine the specific amount of S-type and N-type cost variation that is implicitly cross-subsidized across subgroups in case of premium rate regulation.

In Chapter Eight, it is shown that for most of the subgroups defined by self-reported prior medical utilization and self-reported health status, diseases and conditions, the cost variation that is cross-subsidized under community-rating is largely caused by S-type risk factors. However, for the subgroups that can be defined by the number of years that survey respondents belong to the 25% of insured people with the highest total expenses within each year prior to 2002, the predictable profits and losses appear to be mainly caused by N-type risk factors. It follows that for some subgroups of enrollees, premium rate restrictions create implicit cross-subsidies which are largely in accordance with the policy goals of the sponsor; however, for other subgroups these cross-subsidies are not intended by the sponsor.

Since 2006, Dutch insurers have been allowed to risk-rate premiums across the twelve provinces under the Health Insurance Act. From the results in Chapter Eight

it appears that, given the specification of the 2004 Dutch REF equation, the premium rate variation across the twelve Dutch provinces must be mainly attributed to N-type risk factors. This appears to be even more so if the REF weights are purged from the bias caused by N-type cost variation during the estimation phase. These results justify the decision of the Dutch government to allow premium rates to differ across the twelve Dutch provinces; S-type cost variation appears to be already adequately subsidized by the Dutch REF equation. It should be noted that insured people of only one Dutch insurer are included in the study sample, therefore this research result cannot be treated as representative of all Dutch provinces.

The answer to the third research question of this study is that alternative specifications of the 2004 Dutch REF model can substantially improve the cross-subsidies. Up to 79.5% of the cross-subsidization as intended by the Dutch government can be achieved by application of an ex-post cost sharing arrangement; however, this arrangement reduces the incentives for efficiency at the same time. This drawback does not exist if new adjusters such as PMCGs, MDCGs and MPCGs are added to the 2004 set of REF adjusters; or if a multiplicative instead of additive specification of the REF equation is applied. In these latter cases 77.1% and 75.0% of the effective cross-subsidies can be achieved respectively.

To the extent that the improved cross-subsidies do not yet fully meet the policy goals of the sponsor, premium rate restrictions may induce the necessary cross-subsidization among the subgroups defined by the S-type risk factors. However, for other subgroups of insured people, these rate restrictions may also induce undesired cross-subsidization for N-type cost variation. At the same time, they also create incentives for selection which may have several adverse effects on quality of care, affordability of coverage for high-risk insured people and efficiency in the production of care (Van de Ven, Van Vliet and Lamers 2004). Application of premium rate restrictions as a supplement to risk-adjusted premium subsidies must therefore be necessary and proportional to achieve the stated policy goals; it also demands a careful tradeoff with the incentives for selection (and their possibly adverse effects) which are induced at the same time.

An answer to the central question

The answer to the central question of this study is that the REF adjusters in the 2004 Dutch REF equation generate cross-subsidies up to 71.2% of what is desired given the policy goals of the Dutch government. This achievement can be improved to 79.5% by application of an ex-post risk sharing arrangement, at the expense of the incentives for efficiency. As an alternative strategy to improve the cross-subsidies from the 2004 Dutch REF equation, new adjusters such as

PMCGs, MDCGs and MPCGs appear to be good candidates, and additional research on a multiplicative instead of additive specification of the REF equation is recommended. Premium rate regulation should preferably only hold up to the extent that it is necessary as a supplement to the risk-adjusted premium subsidies in order to meet the specific policy goals of the sponsor; because it creates incentives for selection at the same time. Finally, the use of adjusted REF weights are recommended instead of the original REF weights corresponding to the REF adjusters in any specification of the REF equation, although doing this at the expense of cross-subsidization for S-type cost variation must be avoided as much as possible.

9.2 Discussion

General policy recommendations

The model specifications in Chapters Six, Seven and Eight are used to illustrate the use of the test procedure developed in this study. The theoretical framework developed in this study may be applied to test any risk equalization model that a sponsor has implemented in a competitive health insurance market for the effectiveness of the cross-subsidies they induce in practice. In general, this theoretical framework can also be applied to other sectors where indirect standardization of large populations is guided by a normative decision rule, rather than merely being a statistical exercise.

The first policy recommendation is that the sponsor should always make an explicit choice about the specific categorization of the S-type and the N-type risk factors when implementing a risk equalization model. Given such an explicit choice, the theoretical framework developed in this study can be applied in order to determine the extent to which a REF model safeguards financial access to coverage for high-risk insured people.

The second policy recommendation is to apply the theoretical framework developed in this study on a regular basis to a national sample of insured people to test for the effectiveness of the cross-subsidies, and to check whether the adjusted REF weights should be used instead of the original REF weights in case of imperfect REF adjusters in practice.

The third policy recommendation is to improve cross-subsidies for the insured population with functional impairments by the development of REF adjusters based on utilization of physiotherapy, medical devices and pharmaceutical drugs for mental diseases. The latter category of REF adjusters may prove especially valuable if – contrary to this study – the cost definition also includes mental health care costs.

The fourth policy recommendation is to include an ex-post cost-sharing arrangement in the REF model as a supplement to an incomplete set of REF adjusters. From this study, it appears that ex-post risk sharing increases financial access to coverage for those at high risk even more than the implementation of additional REF adjusters based on the utilization of physiotherapy, medical devices and pharmaceutical drugs for mental diseases. However, ex-post risk sharing also reduces the incentive for efficiency. Thus, there exists a tradeoff between the effectiveness of the cross-subsidies and the incentives of efficiency with ex-post risk sharing.

The fifth policy recommendation is to allow premiums to be risk rated for cost variation caused by N-type risk factors. Premium rate restrictions may hold with respect to S-type risk factors, given that cost variation caused by S-type risk factors is cross-subsidized by the REF model and/or ex-post risk sharing. However, by definition, the (implicit) cross-subsidies that are induced by the premium rate restrictions are not supposed to compensate for cost variation caused by N-type risk factors. Furthermore, premium rate restrictions create incentives for selection which may have adverse effects on quality, affordability and efficiency. The undesirable compensation for N-type cost variation and (the possibly adverse effects of) the incentives for selection can be avoided if premiums are allowed to be risk-rated for cost variation caused by N-type risk factors.

If the sponsor desires cross-subsidies for cost variation caused by the S-type risk factors alone – for example, age, gender and health status – then a proportional implementation of premium rate restrictions would be a ban on the rating factors that are measures of these S-type risk factors (including the REF adjusters), but not on all other rating factors that are measures of the N-type risk factors as is done when requiring community rating. A less stringent alternative to community rating per insurer, per product might be rate-banding. Given the opportunity of rate-banding, insurers will then indicate which relevant measures of the S-type risk factors should actually be added to the REF equation in the next few years. Community rating per premium risk group or class can be implemented by the sponsor in order to protect the consumers against too strong premium increases in case of new rating factors. In any case, the question remains: why does the Dutch government still forbid insurers to rate their premiums based on measures of N-type risk factors, such as, regional input prices, practice style and propensity for consumption?

Country-specific policy recommendations

In the Netherlands, under the 2006 Health Insurance Act, government is legally obliged to undertake a scientific evaluation of the risk equalization system by a panel of international experts in 2008 and 2011 (MoHWS 2005, page 26). The theoretical

framework developed in this study can be applied to test for the effectiveness of the cross-subsidies given a stratified sample of the total Dutch population for this purpose. Such a sample can be found in the Permanent Survey of Living Conditions (POLS) that is conducted yearly by the Dutch National Bureau of Statistics (CBS). The 'Health Status' module in this survey is largely similar to the 2001 Agis Health Survey, and can be applied to derive normative costs for a national study sample of insured people (the SF-12 is used instead of the SF-36).[152] In 2008, it is expected that curative mental health care will be included in the basic benefits package, and the relative importance of the four mental SF-36 scales in the normative equation is expected to differ from that presented in this study.

In Switzerland, the REF adjusters are age, gender and the canton in which an enrollee lives, whereas ex-post risk sharing is not applied as a supplement (Beck et al. 2003, Van de Ven et al. 2007). In 1996 it was decided by law that the model specification would remain unaltered for a period of 10 years, yet in 2004 the law changed, and the risk equalization system was prolonged until 2010 (Bundesrat 2004, Bundesrat 2005). In 2006 the first chamber voted in favor of legislation to include prior hospitalization in hospitals and nursing homes (if admitted for at least three days) as a REF adjuster in the Swiss REF equation (Ständerat 2006, Art. 18a, Abs. 2, page 76). Furthermore, risk equalization is accepted as being permanent. Additional (yet unknown) health indicators are accepted as a long run option and the REF weights are going to be calculated prospectively instead of retrospectively in the future. The National Council did not make a decision after a hearing in May 2006; however, the National Council is expected to vote in favor of this legislation in 2007. Before the end of 2010, the Swiss government must make up their minds with respect to the inclusion of other morbidity-related risk adjusters in the REF equation (for example, PCGs), the implementation of a high-risk sharing arrangement or even the option to abolish the risk equalization system entirely (BBI 2004 4259, page 4273). The theoretical framework developed in this study can be of value in this delicate Swiss debate, as it clearly defines the extent to which cross-subsidies are needed to reduce the market mechanism problems that the Swiss regulators face (Van de Ven et al. 2007). Appendix A6.1 and Section 7.2 show the consequences of these decisions based on the Dutch sample used in this study. Note that the Swiss definition of the benefits package also includes (nursing) home care. It is expected that the specifications of the Dutch REF model

152. The regular frequency to repeat the test procedure should be bi-annually at least, because of sample size limitations that hold for the national survey. The net response to the 'Health Status' module in the yearly POLS survey is targeted to be about 10,000 respondents, whereas the 2001 Agis Health Survey contains 18,617 records. In general, the recommended size depends on the level of detail which is necessary to define the REF equation and the normative equation in this context.

presented in Chapters Six and Seven will induce less effective cross-subsidies if (home) care is included in the Dutch benefits package as well. Therefore, it is expected that the challenge to find an adequate specification will turn out to be at least as large as that for the Dutch REF model.

In Germany, the implementation of morbidity-related REF adjusters to capture the S-type risk factor health status in the REF equation was originally planned in 2007, but is now scheduled for 2009 and is expected to capture 50-80 conditions (Büchner and Wasem 2003, Bundesrat 2007). There will also be a move from an internal modality to channel the REF payments to an external modality, i.e. to modality B in Van de Ven et al. (2000, p. 324). In 2002 risk sharing was introduced for 60% of the costs above a threshold of more than 20,000 euros, thereby increasing the level of ex-post cost compensation from 0% to 4% (in 2006) (Van de Ven et al. 2007, Table 1). In 2003 the (voluntary registration for an) accredited so-called Disease-Management-Program (DMP) has been added as a REF adjuster to the German REF equation. Both the ex-post risk sharing and the DMPs are seen as temporary measures and will be abolished with the planned introduction of a risk equalization fund in 2009. In 2004, a group of international experts advised to use PCGs (RxGroups) and DCGs (HCC) for this purpose (IGES/Lauterbach/Wasem 2004); in 2006, the debate on the implementation of morbidity related REF adjusters was still ongoing (Schokkaert et al. 2006, Table 1). The two-tier insurance system for social and private health insurance will be maintained after 2009; however, sickness funds will be allowed to charge higher premiums for their members than the nation-wide premium based on expectations by the German government. Private health insurers will be confronted by open enrollment regulation for a specified benefits package that is equal to that for sickness funds. Furthermore, risk loading is not allowed when setting premiums. And the premium rates will also be capped, where the cap depends on the average premium rate in the sickness fund sector. There will be some form of risk pooling as a consequence of this premium rate regulation; however, the model specification is yet unknown. The effect of the PCGs and DCGs in the sickness fund sector and the effect of (the implementation of) a system of ex-post risk sharing in both social and private health insurance can be determined by application of the test procedure that is developed in this study.

In Israel, age is the only REF adjuster included in the REF equation. As of 2005, there are eleven instead of nine age subgroups. Gender cannot (yet) be made available because of feasibility problems and no ex-post risk sharing scheme is implemented (Shmueli, Chernichovsky and Zmora 2003). There has been a growing dissatisfaction with this formula. It is argued that children are overcompensated, while the elderly are under compensated, and arguments have been put forward

to include more risk adjusters (Van de Ven et al. 2007). In 2006, the debate on the implementation of morbidity related REF adjusters was still ongoing (Schokkaert et al. 2006, Table 1). Ability to work is not included as a REF adjuster because this is not a relevant measure of health status. Furthermore, socio-economic characteristics are not included as REF adjusters under the hypothesis that time price ("time to visit a doctor") is a more important determinant of observed cost variation than health status. However, based on the results presented in Table 6.7 it must be concluded that there is no good reason not to include ability to work or specific socio-economic characteristics in the Israeli REF equation, given that an adjustment of the REF weights can be applied to take care of cost variation caused by N-type risk factors such as the time price of self-employed insured people. In 2005 it was agreed that the Israeli REF equation would be updated after every four years, therefore a potential revision of their REF equation may be due in 2009. In the meantime, the theoretical framework developed in this study can be applied in Israel to determine the research results that apply to their specific situation. It should be noted that the Israeli formula is not based on individual claims data supplied by the sickness funds, but on a health services utilization survey and a hospitalizations data collection. Data on costs of drugs are not included at all.

In Belgium, the estimation of the REF weights has been based on individual data since 2002. The Belgian REF adjusters are age, sex, morbidity related and socio-economic variables; for example, indicators of chronic illness and disability categories. The information on DRGs and on pharmaceutical consumption is being collected but has not yet been implemented. However, there is a general consensus about the desirability to include it in the future risk adjustment model (Schokkaert et al. 2006). In the on-going debate about the categorization of the S-type and N-type risk factors, many observers keep arguing in favor of a risk adjustment formula which includes as many variables as possible, for example "number of days in the hospital" which is presented as an indicator of morbidity. It was decided not to include medical supply in the REF equation. This held sickness funds responsible for the regional cost variation that is caused by medical supply, although sickness funds do not have adequate instruments to influence the expenditures of their members (Van de Ven et al. 2007). The theoretical framework developed in this study can be used to evaluate the decision with respect to regional cost variation in the Belgian setting.

Since 2004, CMS in the USA has applied a "frailty" REF adjuster to finance PACE organizations. In the context of Medicare, this serves community populations with functional impairments for integrated care so that they can live in their own homes instead of being institutionalized. Due to several methodological problems of feasibility associated with the use of survey data for calculating risk equalization

payments, a "frailty" REF adjuster shall not be implemented program-wide in the CMS-HCC REF equation for Medicare Advantage plans in 2008 (CMS 2007, Attachment II, Section A). However, CMS announces that it will continue to explore ways to incorporate factors into the CMS-HCC REF equation that will better predict costs associated with the "frailty" of individual beneficiaries. From the results presented in this study, utilization of physiotherapy and medical devices may be used to overcome the feasibility problems that the CMS is currently faced with. In the meantime, the REF weights associated with the current REF adjusters can be adjusted in order for the cross-subsidies to better capture cost variation due to functional impairments.

Limitations of this study

The study sample used to illustrate the application of the theoretical framework developed in this study consists of those enrolled with Agis sickness fund in both 2001 and 2002. The results presented in this study may therefore not be representative of all Dutch regions, nor are they representative of the formerly private health insurance population which has also been insured under the Dutch Health Insurance Act since 2006. The cross-subsidies from the 2004 Dutch REF equation are not calculated for insured people younger than 16 years of age in this study, because the health survey was not conducted under this population.

The specification of the 2004 Dutch REF equation differs from its implementation in this study in some respects. The REF adjuster age defines ten-year instead of five-year classes of age, and interactions between the REF adjusters insurance eligibility and age are absent in this study. The cost definition includes production-independent hospital costs in this study; however, these costs are 95% reimbursed in Dutch practice up to 2006. Since 2006, about one third of hospital costs are defined as production-independent; they are effectively 100% reimbursed. Lastly, the proportional risk sharing (amongst insurers) and retrospective reimbursement scheme (between an individual insurer and the sponsor) which existed in 2004 with respect to production-dependent and medical specialty costs, is not applied in this study.

It may be possible that the S-type adjusters included in the normative equation in this study do not capture cost variation caused by S-type risk factors to the full extent. The range of health indicators may be expanded depending upon availability in future studies. The crux of the application of the theoretical framework in this study is that the current array of S-type adjusters is less limited than the set of REF adjusters which are used in practice. In this sense, the approach developed in this study produces a lower bound on the extent to which the REF equation induces the cross-subsidies that the sponsor desires. In other words, given the implementation of normative costs in this study, the performance scores of the REF models as

presented in Chapters Six, Seven and Eight will probably indicate a maximum for the extent to which the policy goals of the Dutch government are met.

The omitted variables approach to removing the bias from the imperfect REF weights appears to be rather limited given the set of N-type adjusters applied in this study. Results might change if another potentially broader set of N-type adjusters are applied.

Further research

The empirical results presented in this study apply to the 2002 Agis sickness fund insured people. Although the REF weights are not expected to change significantly for most REF adjusters, the exercise should be repeated for the total Dutch population of sickness fund enrollees, in order to be representative in this respect. In particular, the relative importance of the regional REF weights might change as a result.

Furthermore, the Dutch REF model holds with respect to all 16 million Dutch citizens under the 2006 Dutch Health Insurance Act. Therefore, the theoretical framework developed in this study is relevant to the total Dutch population since then and should be applied as such. The normative equation can be implemented based on information from national surveys which are currently available; for example, the 'Health Status' module in the POLS survey of the CBS.

Within a competitive health insurance market without risk equalization, premium rebates under the option of voluntary deductibles will reflect cost variation caused by S-type risk factors as a consequence of adverse selection (Van Kleef, Van de Ven and Van Vliet 2006). To some extent, this will still be the case if cross-subsidies are based on incomplete and/or imperfect REF adjusters. The extent to which this is the case can be determined under the approach developed in this study. Furthermore, the level of the voluntary deductible chosen by the insured people may be included in the REF equation as a proxy for health status. Although implementation of this proxy as a new REF adjuster may also induce undesired cross-subsidies for N-type cost variation, these effects can be explicitly weighed against each other by application of the theoretical framework developed in this study. An adjustment of the corresponding REF weight can be applied in order to avoid compensation for these N-type effects. Note that an analogous exercise is already performed with respect to insurance eligibility and region in the empirical part of this study.

The ultimate goal of managed competition is that insurers take up their role as prudent purchasers of health care. For this situation to occur, consumers must not only switch insurers as a consequence of their sensitivity to observed premium differences among insurers, but consumers must also be enabled to observe and

be sensitive to quality differences of health care delivery. Under the pressure of market competition, insurers are then forced to organize and purchase/provide health care according to the preferences of their enrollees. A more direct way to stimulate insurers to meet consumers' preferences, is to let the cross-subsidies depend on explicit measures of the quality of the health care that they contracted (in addition to compensation for S-type cost variation). The IOM (2006) recommends creating pools from a reduction in the base Medicare payments for each class of providers (hospitals, skilled nursing facilities, Medicare Advantage plans, dialysis facilities, home health agencies, and physicians). Initially, pay-for-performance programs should be designed such that providers are rewarded who achieve high performance and manage to improve performance significantly over time. Given that these indicators are often available for only a limited number of consumers, a natural approach seems to be to include these measures in the normative equation such that these are reflected in the adjusted REF weights.

A well-documented case of under provision necessary health care of social subgroups can be found in Shmueli (2000). This is for Arab insured people living in Israel. Table A8.3 of this study shows that REF predicted costs for first-generation immigrants are slightly above observed costs, but at the same time it is revealed that REF predicted costs are significantly lower than normative costs. This means that there is severe underutilization by first-generation immigrants. In this case there is no financial incentive for an insurer to tackle this problem of underutilization, because REF predicted costs are barely different from actual costs. This generates incentives for adverse selection against first-generation immigrants, yet removing the problem will lead to predictable losses for the subgroup of first-generation immigrants. Although this problem should probably be targeted by direct subsidies or educational programs, according to Schokkaert et al. (2006), an alternative approach may be to include first-generation immigrants as a REF adjuster in the REF equation and adjust the corresponding REF weight for N-type cost variation at the same time. This approach makes them preferred risks for insurers. Note however, that insurers must then also be allowed to adjust their premiums for the subgroup of first-generation immigrants in order to reduce the danger that the resulting predictable profits may be allocated for other purposes than combating underutilization.

In sum, the approach to risk equalization developed in this study can be applied in practice in several ways, and it is relevant to all countries with competitive health insurance markets. This approach is recommended to test for and improve the cross-subsidies of REF models in these cases.

Gloss.

GLOSSARY

Term	Definition
Acceptable costs	Acceptable costs are defined as the costs associated with the level of intensity, quality and (demand and supply) price of treatment that the sponsor has decided to be acceptable to be subsidized. Acceptable costs are thus defined at the individual level of treatment. The sponsor may decide which levels are acceptable and which are not, for example, those generated if only medically necessary and cost-effective care is provided. Thus, in general, acceptable costs deviates from observed costs.
Adverse selection	Adverse selection is the selection that occurs because high-risk consumers have an incentive to buy more coverage than low-risk consumers within the same premium risk group. These actions by consumers may arise if consumers have an information surplus over the insurers, which may be the result of the government regulation (i.e. premium rate restrictions) on the health insurance market or because of asymmetric information between consumers and insurers which may even exist in unregulated competitive health insurance markets (Wilson 1977).
Affordability	Affordability of coverage for high-risk insured people is achieved if the premium subsidies are such that they are able to pay their premiums in a competitive individual health insurance market. Note that the goal of a sponsor is to make premiums affordable up to the extent that their rates are determined by cost variation caused by S-type risk factors alone.
Conventional risk adjustment	The estimation of a linear model containing demographic and/or previous utilization variables to predict actual expenditures at the level of individual enrollees. In practice, only a limited range of potential risk adjusters is available and an explicit decision is made about which specific variables are measures of S-type risk factors to be included in the REF equation and which are not. In most countries, measures of the N-type risk factors are omitted during the estimation stage.
Cross-subsidies	Subsidies between high-risk and low-risk enrollees, induced by the REF model. See premium subsidy.
Efficiency	Technical efficiency or so-called efficiency in the production of care, not allocative efficiency.
Imperfect REF adjusters	REF adjusters are called imperfect if they capture cost variation caused by N-type risk factors which is reflected as biased REF weights.
Incomplete REF adjusters	REF adjusters are called incomplete if they do not fully capture the cost variation that is induced by the S-type risk factors.
Needs	Needs for health care are unobserved. In resource allocation formula needs are expressed by measures of health and socio-economic variables up to the extent that they reflect needs.
Norm, Normative	(1) Norms are ways of behaving that are considered normal in a particular society; (2) If you say that a situation is the norm, you mean it is usual and expected; (3) A norm is an official standard or level that organizations are expected to reach. Normative means creating or stating particular rules of behavior (Sinclair 2001). The norm in this study is that cost variation should be subsidized insofar as being caused by S-type risk factors only.

Term	Definition
Normative costs	Normative costs are based on actual expenditures for subgroups of insured people, where the subgroups are defined by the S-type adjusters as represented by the independent variables X given in equation (2.4). Normative costs are often defined as the statistical average of observed costs at the subgroup level, also in this study, as denoted by the dependent variable Y^{NORM} in equation (2.4).
Open enrollment	A periodic open enrollment requirement implies that during the open enrollment period, for example one month every year, consumers are allowed to change insurer and each insurer must accept anyone who wants to join.
Overcompensation	REF predicted costs are larger than acceptable costs on average for some subgroup of insured people.
Overutilization	Acceptable costs are smaller than observed costs on average for some subgroup of insured people.
Preferred selection	Preferred selection is the selection that occurs because insurers prefer low-risk consumers to high-risk consumers within the same premium risk group (Van de Ven and Ellis 2000). Preferred selection (or "cream skimming", or "cherry picking") may be undertaken by insurers in regulated health insurance markets if they have an information surplus over the Risk Equalization Fund. There are financial incentives for cream skimming if the profits of these actions outweigh the costs of this behavior.
Premium subsidy	In theory, the risk-adjusted premium subsidies mentioned in this study are a function of acceptable costs of individual enrollees. In practice, they are a function of the average predicted per capita expenditures for the subgroup defined by the REF adjusters to which the beneficiary belongs, e.g. a percentage function or simply a fixed amount is subtracted.
REF	The term REF is an abbreviation of Risk Equalization Fund and is used instead of the term sponsor as used by Van de Ven and Ellis (2000) to highlight the redistribution role, in addition to the fact that the REF may also regulate the characteristics of health plans that are offered (Newhouse 1996). In general, a REF can be an employer, a coalition of employers, a government agency, a nonprofit organization, or a distinct insurance entity empowered to use coercion to redistribute risk.
Regulator	See REF.
Risk equalization	The procedure to equalize cost differences among subgroups of insured people as organized by a sponsor among health insurers.
Risk factor	Van de Ven and Ellis (2000) distinguish seven types of risk factors that may explain structural variation in observed costs: age/gender, health status, socio-economic characteristics, provider characteristics, input prices, market power of the insurer and benefit plan characteristics, see also Figure 2.1.
Risk selection	See preferred selection.
Risk sharing	Risk sharing implies that the insurers are retrospectively reimbursed by the sponsor for some of the costs of some of their insurance members (Van de Ven and Ellis 2000). Consequently the risk-adjusted premium subsidies have to be adjusted to the insurers' new financial risk.

Term	Definition
Selection	Actions (not including premium differentiation) by consumers and insurers to exploit unpriced risk heterogeneity and break pooling arrangements (Newhouse 1996). The literature identifies two forms of selection: adverse selection and cream skimming. These forms of selection are different from each other in terms of the type of selection actions that may actually be undertaken by consumers and insurers, as well as in their effects on efficiency and solidarity.
Sickness fund	Until 2006, sickness funds are Dutch health plans that purchase health care for their members under a social health insurance scheme. Since 2006, all Dutch citizens have contracts with private health insurers for mandatory social health insurance.
Sponsor	A sponsor reallocates the burden of health insurance premiums across insured people, and enters into risk-sharing arrangements with insurers (Van de Ven and Ellis 2000). In addition, the sponsor may also regulate the characteristics of insurance policies that are offered (Newhouse 1996). In general, a sponsor can be an employer, a coalition of employers, a government agency, a nonprofit organization, or a distinct insurance entity empowered to use coercion to redistribute risk. In many countries, the sponsor role is fulfilled by the government agency that regulates access to individual (or small group) private health insurance coverage in a competitive market. In the US, the role of sponsor is also fulfilled by (large) employers who offer group health insurance to their employees.
Undercompensation	REF predicted costs are smaller than acceptable costs on average for some subgroup of insured people.
Underutilization	Acceptable costs are larger than observed costs on average for some subgroups of insured people.
WOVM	WOVM is an abbreviation of "Werkgroep Ontwikkeling VerdeelModel" (EN: Working Group on the Development of the Risk Equalization Model). In the WOVM working group econometric research on the Dutch risk equalization scheme is monitored and validated under the authority of the Dutch Ministry of Health, Well-Being and Sports (MoHWS). In the WOVM working group there are representatives of MoHWS, a government agency on health insurance (CVZ), the Association of Health insurers (ZN) and health insurers. Each year, the WOVM working group advises the Minister of Health, Well-Being and Sports on the REF risk equalization formula for the then coming year. In 2005, the WOVM working group was renamed to WOR, i.e. "Werkgroep Onderzoek Risicoverevening" (EN: Working Group Research on Risk Equalization). The so-called WOVM (or: WOR) databases that are input to this type of research are constructed by the insurers according to a standardized format and consist of all claims at the individual member level.
WOR	See WOVM.

Abbr.

Abbreviations

Abbreviations	Description
BP	Bodily Pain
DCG	Diagnostic Cost Group
DxG	Diagnosis Group
GH	General Health
MCS	Mental Component Scale
MH	Mental Health
PCG	Pharmacy-Cost Group
PCS	Physical Component Scale
pipy	Per insured, per year
PF	Physical Functioning
RE	Role-Emotional
REF	Risk Equalization Fund
RP	Role-Physical
SF	Social Functioning
VT	Vitality

Refs.

References

Aaronson, N.K., M. Muller, P.D.A. Cohen, M.L. Essink-Bot, M. Fekkes, R. Sanderman, M.A.G. Sprangers, A. te Velde and E. Verrips (1998), "Translation, Validation, and Norming of the Dutch Language Version of the SF-36 Health Survey in Community and Chronic Disease Populations", Journal of Clinical Epidemiology, 51, 11, 1055–1068.

AD (2003), "Ziekenhuis Top 100" (EN: "The AD ziekenhuis Top 100"), Supplement Leefwereld/Diagnose, 16-17, 13 October 2004, Rotterdam, The Netherlands. See also: http://www.ad.nl/ziekenhuistop100/.

Ash, A., F. Porell, L. Gruenberg, E. Sawitz, and A. Beiser (1989), "Adjusting Medicare premium subsidies using prior hospitalization data", Health Care Financing Review, 10, 4, 17-29.

Ash, A.S., R.P. Ellis, G.C. Pope, J.Z. Ayanian, D.W. Bates, H. Burstin, L.I. Iezzoni, E. MacKay, Wei Yu (2000), "Using diagnoses to describe populations and predict costs", Health Care Financing Review, 21, 3, 7-28.

Basant, E. (2003), "Oordeel klanten bepaalt mede salaris huisarts" ("Customer opinions too determine salary of general practitioner"), Het Financieele Dagblad, June 18, The Netherlands.

Basu, A., W.G. Manning and J. Mullahy (2004), "Comparing alternative models: log vs Cox proportional hazard?", Health Economics, 13, 749-765

Beck, K., S. Spycher, A. Holly, and L. Gardiol (2003), "Risk adjustment in Switzerland", Health Policy, 65, 63-74.

Blough, D.K., C.W. Madden and M.C. Hornbrook (1999), "Modeling risk using generalized linear models", Journal of Health Economics, 18, 153-171.

Botterweck, A., F. Frenken, S. Janssen, L. Rozendaal, M. de Vree, and F. Otten (2003), "Plausibiliteit nieuwe metingen algemene gezondheid en leefstijlen 2001", (EN: "Plausibility of new generic health and lifestyles measurements 2001"), H 539-03-SAH, CBS Netherlands Statistics, Heerlen, The Netherlands.

Box, G., and D. Cox (1964), "An analysis of transformations", Journal of the Royal Statistical Society, Series B, 211-264.

Büchner, F. and J. Wasem (2003), "Needs for further improvement: risk adjustment in the German health insurance system", Health Policy, 65, 21-35.

Bundesrat (2004), "Botschaft zur Änderung des Bundesgesetzes über die Krankenversicherung (Strategie und dringliche Punkte)", Bundesblatt, 29, 4259-4288, Bern, Switzerland.

Bundesrat (2005), "Bundesgesetz über die Krankenversicherung (KVG) (Gesamtstrategie und Risikoausgleich): Änderung vom 8. Oktober 2004", Amtliche Sammlung des Bundesrechts, 1071-1074, Bern, Switzerland.

Bundesrat (2007), "Gesetz zur Stärkung des Wettbewerbs in der gesetzlichen Krankenversicherung (GKV-Wettbewerbsstärkungsgesetz – GKV-WSG)" (EN: A law to increase competition in social health insurance), 75/07, 1-127, Bundesanzeiger Verlasgesellschaft mbH, Köln, Germany.

CAHPS® (2002), "Article 9: Determining a complete questionnaire", Document No. 109, 10/01/02, CAHPS® Survey and Reporting Kit 2002, AHRQ, Department of Health and Human Services, Rockville MD, USA.

Campbell, D.T. and D.W. Fiske (1959), "Convergent and discriminant validation by the multitrait-multimethod matrix", Psychological Bulletin, 56, 81-105.

Carr-Hill, R.A., T.A. Sheldon, P. Smith, S. Martin, S. Peacock, and G. Hardman (1994), "Allocating resources to health authorities: development of method for small area analysis of use of inpatient services", British Medical Journal, 309, 1046-1049.

CBS (2004), "Statistisch Jaarboek 2004", (EN: "Statistical Yearbook 2004"), CBS Netherlands Statistics, Voorburg / Heerlen, The Netherlands.

Clark, D.O., M. von Korff, K. Saunders, W.M. Baluch and G.E. Simon (1995), "A chronic disease score with empirically derived weights", Medical Care, 33, 783-95.

CMS (2007), "2008 Advance Notice of Methodological Changes for Calendar Year (CY) 2008 Medicare Advantage (MA) Capitation Rates and Part D Payment", http://www.cms.hhs.gov/MedicareAdvtgSpecRateStats/Downloads/Advance2008.pdf, Last visited: February 26th, 2007.

Cronbach, L. (1951), "Coefficient alpha and the internal structure of tests", Psychometrika, 16, 3, 297–334.

CVZ (2002), "CVZorgcijfers 1997-2001" (EN: "CVZ Health Care Figures 1997-2001"), CVZ, Amstelveen, The Netherlands.

CVZ (2003), "Herziening verpleegdagtarieven: gemiddelde verpleegtarieven 2002" (EN: "Revision nursing day tariffs: average hospital tariffs 2002"), Supplement, VFIN23005616, January 31st, CVZ, Amstelveen, The Netherlands.

Den Dulk, C. J., H. van der Stadt, and J. M. Vliegen (1992). "Een nieuwe maatstaf voor stedelijkheid: de omgevingsadressendichtheid" (EN: A new criterion for urbanization: address density of the surrounding area), Maandstatistiek van de bevolking, July, 14-27, CBS Netherlands Statistics, Heerlen, The Netherlands.

Duan, N. (1983), "Smearing estimate: A nonparamteric retransformation method", Journal of the American Statistical Association, 78, 383, 605-610.

Ellis, R.P., and A. Ash (1995), "Refinements to the diagnostic cost group model", Inquiry, 32, 4, 418-429.

Ellis, R.P., G.C. Pope, L.I. Iezzoni, et al. (1996), "Diagnosis-based risk adjustment for Medicare premium subsidies", Health Care Financing Review, 17, 3, 101-128.

Enthoven, A.C. (1978), "Consumer choice health plan", New England Journal of Medicine, 298, 650–658 and 709–720.

Enthoven, A.C. (1988), "The Theory and Practice of Managed Competition", Amsterdam, North-Holland.

G.C. Pope, R.P. Ellis, A.S. Ash, C.F. Liu, J.Z. Ayanian, D.W. Bates, H. Burstin, L.I. Iezzoni, M.J. Ingber (1999), "Principal inpatient diagnostic cost group models for Medicare risk adjustment", final report prepared for Health Care Financing Administration, Health Economics Research, Inc. Waltham, MA.

Garratt, A.M., L. Schmidt, A. Mackintosh, R. Fitzpatrick (2002), "Quality of life measurement: bibliographic study of patient assessed health outcome measures", British Medical Journal, 324, 1417-1421.

Gravelle, H., M. Sutton, S. Morris, F. Windmeijer, A. Leyland, C. Dibben and M. Muirhead (2003), "Modelling supply and demand influences on the use of health care: implications for deriving a needs-based capitation formula, Health Economics, 12, 985-1004.

Gruenberg, L., C. Tompkins and F. Porell (1989), "The health status and utilization patterns of the elderly: implications for setting Medicare payments to HMO's", in

R.M. Scheffler and L.F. Rossiter (eds.), Advances in health economics and health services research, JAI Press, Greenwich, 10, 41-73.

Haffer, S.C., S.E. Bowen, E.D. Shannon, and B.M. Fowler (2003), "Assessing Beneficiary Health Outcomes and Disease Management Initiatives in Medicare", Disease Management and Health Outcomes, 11, 111-124.

HEDIS® (2003), "Volume 6: Specifications for the Medicare Health Outcomes Survey", NCQA, Washington DC, USA.

Hornbrook, M.C. and M.J. Goodman (1995), "Assessing relative health plan risk with the Rand-36 health survey", Inquiry, 32, 56-74.

Hornbrook, M.C. and M.J. Goodman (1996), "Chronic Disease, Functional Health Status, and Demographics: A Multi-Dimensional Approach to Risk Adjustment", Health Services Research, 31, 1, 283-307.

Iezzoni, L.I. (2003), "Risk Adjustment for Measuring Health Care Outcomes", 3rd edition, Health Administration Press, Chicago, Illinois, USA.

Iezzoni, L.I., E.P. McCarthy, R.B. Davis and H. Siebens (2001), "Mobility Difficulties Are Not Only a Problem of Old Age", Journal of General Internal Medicine, 16, 4, 235-43.

IGES/Lauterbach/Wasem (2004), "Klassifikationsmodelle für Versicherte im Risikostrukturausgleich. Untersuchung zur Auswahl geeigneter Gruppenbildungen, Gewichtungsfaktoren und Klassifikationsmerkmale für einen direkt morbiditätsorientierten Risikostrukturausgleich in der gesetzlichen Krankenversicherung", Berlin/Köln/Essen, Germany.

Jones, N., S.L. Jones, and N.A. Miller (2004), "The Medicare Health Outcomes Survey program: Overview, context, and near-term prospects", Health and Quality of Life Outcomes, 2, 33.

Kautter, J. and G.C. Pope (2005), "CMS Frailty Adjustment Model", Health Care Financing Review, 26, 2, 1-19.

Lamers, L. (2000), "Predictive power of survey variables for health care expenses in the Z&Z survey", iBMG, Erasmus University Rotterdam, The Netherlands, unpublished.

Lamers, L.M. (1995), "Gezondheidsenquête onder verzekerden van zorgverzekeraar Zorg en Zekerheid: Een beschrijvende analyse", (EN: "Health Survey conducted under members of health insurer Zorg en Zekerheid: A descriptive analysis"), iBMG, Erasmus University Rotterdam, The Netherlands.

Lamers, L.M. (1997), "Capitation payments to competing Dutch sickness funds based on diagnostic information from prior hospitalizations", Ph.D. thesis, Ridderprint, Ridderkerk, The Netherlands.

Lamers, L.M. (1997), "Premium subsidies to competing Dutch sickness funds based on diagnostic information from prior hospitalizations", Ph.D. thesis, Ridderprint, Ridderkerk, The Netherlands.

Lamers, L.M. (1998), Risk-adjusted premium subsidies: developing a diagnostic cost groups classification for the Dutch situation, Health Policy 45, 15-32.

Lamers, L.M. (1999a), "Pharmacy Costs Groups: a risk-adjuster for premium subsidies based on the use of prescribed drugs", Medical Care, 37, 8, 824–30.

Lamers, L.M. (1999b), "Risk-adjusted capitation based on the Diagnostic Cost Group model: An empirical evaluation with health survey information", Health Services Research, 33, 6, 1727-1744.

Lamers, L.M. and R.C.J.A. van Vliet (1996), "Multiyear diagnostic information from prior hospitalizations as a risk adjuster for premium subsidies", Medical Care, 34, 549-561.

Lamers, L.M., and R.C.J.A. van Vliet (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., and R.C.J.A. van Vliet (2004), "The Pharmacy-based Cost Group model: validating and adjusting the classification of medications for chronic conditions to the Dutch situation", Health Policy, 68, 113,121.

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

Lamers, L.M., R.C.J.A. van Vliet (2004), "The Pharmacy-based Cost Group model: validating and adjusting the classification of medications for chronic conditions to the Dutch situation", Health Policy, 68, 113-121.

Lamers, L.M., R.C.J.A. van Vliet, and W.P.M.M. van de Ven (2003), "Risk adjusted premium subsidies and risk sharing: key elements of the competitive sickness fund market in the Netherlands", Health Policy, 65, 49-62.

Likert, R. (1932), "A technique for the measurement of attitudes", Archives of Psychology, 140, 5-55.

Mackenbach, J.P., C.W.N. Looman and J.B.W. van der Meer (1996), "Differences in the Misreporting of Chronic Conditions, by Level of Education: The Effect on Inequalities in Prevalence Rates", American Journal of Public Health, 86, 5, 706-711.

Manning, W.G. and J. Mullahy (2001), "Estimating log models: to transform or not to transform", Journal of Health Economics, 20, 461-494.

Manning, W.G., A. Basu and J. Mullahy (2003), "Generalized modeling approaches to risk adjustment of skewed outcomes data", NBER Technical Working Paper No. 293.

Manning, W.G., A. Basu and J. Mullahy (2005), "Generalized modeling approaches to risk adjustment of skewed outcomes data", Journal of Health Economics, 24, 3, 465-488.

Marchand, M., M. Sato, and E. Schokkaert (2003), "Prior year expenditures and risk sharing with insurers competing on quality", The RAND Journal of Economics, 34, 4, 647-669.

McHorney, C. A., J.E. Ware, R.L. Lu and D. Sherbourne (1994), "The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups", Medical Care, 32, 1, 40-66.

McHorney, C.A., J.E. Ware, and A.E. Raczek (1993) "The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and Clinical Tests of Validity in Measuring Physical and Mental Health Constructs", Medical Care, 31, 3, 247-263.

MoHWS (2004), "Health Insurance in the Netherlands: Status as of 1 January 2004", International Publication Series Health, Welfare and Sport no. 1E, Ministry of Health, Welfare and Sport, The Hague, The Netherlands.

MoHWS (2005), "Besluit Zorgverzekering" ("Health Insurance Decision"), Staatsblad 389, Ministry of Health, Welfare and Sports, SDU Uitgevers, The Hague, The Netherlands.

National Committee for Quality Assurance (2003), "Medicare Health Outcomes Survey Instrument: Cohort I Performance Measurement Data User's Guide", Washington DC, USA.

National Committee for Quality Assurance (2005), "Narrative: What's in It and Why It Matters", NCQA, HEDIS® 2005, Volume 1, Washington DC, USA.

Newhouse, J.P. (1993), "Free for All? Lessons from the RAND Health Insurance Experiment", Harvard University Press, Cambridge, USA.

Newhouse, J.P. (1996), "Reimbursing health plans and health providers: efficiency in production versus selection", Journal of Economic Literature, 34, 1236-1263.

Nunnally, J. (1978), "Psychometric theory", 2nd edition, New York NY, McGraw-Hill.

Paolucci, F., A. den Exter, and W.P.M.M. van de Ven (2006), "Solidarity in competitive health insurance markets: analysing the relevant EC legal framework", Health Economics, Policy and Law, 1, 107-126.

Pope G.C., J. Kautter, R.P. Ellis, A.S. Ash, J.Z. Ayanian, L.I. Iezzoni, M.J. Ingber, J.M. Levy, J.Robst (2004), "Risk adjustment of Medicare premium subsidies using the CMS-HCC model", Health Care Financing Review, 25, 4, 119-141.

Pope, G.C., Ellis, R.P., Ash, A.S., Chuan-Fen Liu, J.Z. Ayanian, D.W. Bates, H. Burstin, L.I. Iezzoni, and M.J. Ingber (2000), "Principal Inpatient Diagnostic Cost Group Model for Medicare Risk Adjustment", Health Care Financing Review, 21, 3, 93-118.

Pope, G.C., R.P. Ellis, C.F. Liu et al. (1998), "Revised Diagnostic Cost Group (DCG)/ Hierarchical Coexisting Conditions (HCC) Models for Medicare Risk Adjustment", Final Report to the Health Care Financing Administration under Contract Number 500-95-048, Health Economics Research, Inc., Waltham, MA, USA.

Prinsze, F.J., W.P.M.M. van de Ven, D. de Bruijn en F.T. Schut (2005), "Verbetering risicoverevening in de zorgverzekering: van groot belang voor chronisch zieken" ("Improving risk equalization in social health insurance: of great importance to the chronically ill"), institute of Health Policy and Management (iBMG), Erasmus University Rotterdam, The Netherlands.

Ramsey, J.B. (1969), "Tests for Specification Error in Classical Linear Least Squares Regression Analysis", Journal of the Royal Statistical Society, Series B, 31, 350–371.

Ruwaard, D. and P.G.N. Kramers eds. (1997), "De som der delen – Volksgezondheid Toekomst Verkenning 1997" (EN: The sum of the parts - The 1997 Dutch Public Health Status and Forecasts Report), Elsevier/De Tijdstroom, Utrecht, The Netherlands.

SAS Institute Inc. (1999), "SAS OnlineDoc, Version 8", SAS Institute Inc., Cary, NC, USA.

Schauffler, H.H., J. Howland and J. Cobb (1992), "Using chronic disease risk factors to adjust Medicare premium subsidies", Health Care Financing Review, 14, 1, 79-90.

Schokkaert, E., and C. van de Voorde (2000), "Risk adjustment and the fear of markets: the case of Belgium", Health Care Management Science, 3, 121–130.

Schokkaert, E., and C. van de Voorde (2003), "Belgium: risk adjustment and financial responsibility in a centralised system", Health Policy, 65, 5-19

Schokkaert, E., and C. van de Voorde (2004), "Risk selection and the specification of the conventional risk adjustment formula", Journal of Health Economics, 23, 1237-1259.

Schokkaert, E., G. Dhaene and C. van de Voorde (1998), Risk adjustment and the trade-off between efficiency and risk selection: an application of the theory of fair compensation, Health Economics, 7, 465-480.

Schokkaert, E., K. Beck, A. Shmueli,W. van de Ven, C. Van de Voorde, J. Wasem (2006), "Acceptable costs and risk adjustment: policy choices and ethical trade-offs", Social Science Research Network, Working Paper Series, 1-33.

Shmueli, A. (2000), "Inequality in medical care in Israel: Arabs and Jews in the Jerusalem district of the General Sick Fund", European Journal of Public Health 10, 1, 18-23.

Shmueli, A., D. Chernichovsky and I. Zmora (2003), "Risk adjustment and risk sharing: the Israeli experience", Health Policy, 65, 37-48.

Sinclair, J. (Ed.) (2001), "Collins Cobuild English Dictionary for Advanced Learners", 3rd edition, Harper Collins Publishers (2001).

Smith, P.C., N. Rice and R. Carr-Hill (2001), "Capitation funding in the public sector", Journal of the Royal Statistical Society: Series A, 217-257.

Staatsblad (2005), "Besluit Zorgverzekering", 389, SDU Uitgevers, The Hague, The Netherlands.

Ständerat (2006), "Bundesgesetz über die Krankenversicherung. Teilrevision. Spitalfinanzierung", Amtliches Bulletin der Bundesversammlung, 04.061, 70-77, Bern, Switzerland.

StataCorp (2006), "Statistical Software: Release 9.2", College Station, TX, USA, Stata Corporation.

Sutton, M., H. Gravelle, S. Morris, A. Leyland, F. Windmeijer, C. Dibben, and M. Muirhead (2002), "Allocation of resources to English areas: Individual and small area determinants of morbidity and use of health care resources", Report to the Department of Health, Information and Statistics Division, Edinburgh, Scotland, Great Britain.

Thomas, J.W. and R. Lichtenstein (1986), "Including health status in Medicare's adjusted average per capita cost capitation formula", Medical Care, 24, 259-275.

Thomas, R.L. (1985), "Introductory econometrics: theory and applications", Longman, London/New York.

Van Barneveld, E.M. (2000), "Risk sharing as a supplement to imperfect capitation in health insurance: A tradeoff between selection and efficiency", Ph.D. thesis, Ridderprint, Ridderkerk, The Netherlands.

Van de Ven, W.P.M.M. (2001), "Risk selection on the sickness fund market", Health Economics in Prevention and Care, 2, 91-95.

Van de Ven, W.P.M.M. (2006), "Response: The case for risk-based subsidies in public health insurance", Health Economics, Policy and Law, 1, 159-199.

Van de Ven, W.P.M.M., and F.T. Schut (1994), "Should catastrophic risks be included in a regulated competitive health insurance market?", Social Science and Medicine, 39, 10, 1459-1472.

Van de Ven, W.P.M.M., and R.P. Ellis (2000), "Risk Adjustment in Competitive Health Plan Markets", in Handbook of Health Economics, vol. 1, ed. A.J. Culyer and J.P. Newhouse (Amsterdam: Elsevier Science BV, 2000), 755–845.

Van de Ven, W.P.M.M., K. Beck, C. van de Voorde, J. Wasem and I. Zmora (2007), "Risk adjustment and risk selection in Europe: 6 years later", Health Policy, online prepublication, doi:10.1016/j.healthpol.2006.12.004.

Van de Ven, W.P.M.M., K. Beck, F. Buchner, D. Chernichovsky, L. Gardiol, A. Holly, L.M. Lamers, E. Schokkaert, A. Shmueli, S. Spycher, C. van de Voorde, R.C.J.A. van Vliet, J. Wasem, and I. Zmora (2003), "Risk adjustment and risk selection on the sickness fund insurance market in five European countries", Health Policy, 65, 75-98.

Van de Ven, W.P.M.M., R.C.J.A. van Vliet, and L.M. Lamers (2004), "Health- adjusted premium subsidies in the Netherlands", Health Affairs, 23, 3, 45-55.

Van de Ven, W.P.M.M., R.C.J.A. van Vliet, F.T. Schut, and E.M. van Barneveld (2000), "Access to coverage for high-risks in a competitive individual health insurance market: via premium rate restrictions or risk-adjusted premium subsidies?", Journal of Health Economics, 19, 311-339.

Van den Berg, J. and C. van der Wulp (2003), "Rapport van de Werkgroep Revisie POLS-Gezondheidsenquête 1999", (EN: "Report of the Working Group Revision POLS-Health Survey 1999"), H 538-03-SAH, CBS Netherlands Statistics, Heerlen, The Netherlands.

Van den Brink, W.P. and G.J. Mellenbergh (1998), "Testleer en testconstructie" (EN: "Test theory and test construction"), Boom, Amsterdam, The Netherlands.

Van der Meer, J.B.W., J. van den Bos, C.W.N. Looman and J.P. Mackenbach (1996), "Een zorg minder? De longitudinale studie naar sociaal-economische verschillen in medische consumptie (LS-SEVM)", instituut Maatschappelijke Gezondheidszorg, Erasmus Universiteit Rotterdam, Rotterdam.

Van der Zee, K.I., R. Sanderman, and J. Heyink (1996), "A comparison of two multidimensional measures of health status: The Nottingham Health Profile and the RAND 36-Item Health Survey 1.0", 5, 1, 165–174.

Van Kleef, R.C., W.P.M.M. van de Ven and R.C.J.A. van Vliet (2006), "A voluntary deductible in social health insurance with risk equalization: Community-rated or risk-rated premium rebate?", The Journal of Risk and Insurance, 73, 3, 529-550.

Van Oers, J.A.M. ed. (2002), "Gezondheid op koers? - Volksgezondheid Toekomst Verkenning 2002" (EN: Health on course? - The 2002 Dutch Public Health Status and Forecasts Report), Bohn Stafleu Van Loghum, Houten, The Netherlands.

Van Sonsbeek, J.L.A, and L.H. Stronkhorst (1983), "Vergelijking van drie waarnemingsvarianten bij de meting van medische consumptie", (EN: "A comparison of data collection methods in the measurement of medical consumption"), CBS Netherlands Statistics, The Hague, The Netherlands.

Van Vliet, R.C.J.A. (1999), "Alternatieve vormgevingen van het ZFW-verdeelmodel" (EN: "Alternative specifications of the risk equalization model for sickness funds"), WOVM 290, iBMG, Erasmus University Rotterdam, The Netherlands.

Van Vliet, R.C.J.A. (2002), "Opsporen van chronisch zieken op basis van kostenpatronen" (EN: Identifying chronic diseases from patterns in health care expenses), WOVM 406, Erasmus University Rotterdam, The Netherlands.

Van Vliet, R.C.J.A., and L.M. Lamers (2000), "Verdeelkenmerken voor het ZFW verdeelmodel gebaseerd op chronische aandoeningen afgeleid uit medicijngebruik uit het verleden" (EN: "REF adjusters for the sickness fund risk equalization model based on chronic conditions derived from prior utilization of pharmaceutical drugs"), WOVM 370, iBMG, Erasmus University Rotterdam, The Netherlands.

Van Vliet, R.C.J.A., and F.J. Prinsze (2003), "Eindrapportage: Onderhoud FKG's en nader vervolgonderzoek naar DKG's voor toepassing in het ZFW-verdeelmodel 2004" (EN: Final report: Maintenance of PCGs and ensuing research on the applicability of DCGs in the 2004 sickness fund risk equalization model), WOVM 612, Erasmus University Rotterdam, The Netherlands.

Van Vliet, R.C.J.A., R. Goudriaan, and V. Thio (2003), "Overall Toets ZFW-verdeelmodel 2004" (EN: "Overall-Test Sickness Fund RACP model 2004"), WOVM 613, Aarts De Jong Wilms Goudriaan Public Economics bv (APE), The Hague, The Netherlands.

Vektis (2004), "Actualisatie parameters C-voorziening" (EN: Actualization parameter values C-provision), 04-1201, Zeist, The Netherlands

Von Korff, M., E.H. Wagner, and K. Saunders (1992), "A chronic disease score from automated pharmacy data", Journal of Clinical Epidemiology, 45, 197-203.

Ward, J.H. (1963), "Hierarchical Grouping to Optimize an Objective Function", Journal of the American Statistical Association, 58, 236 -244.

Ware, J.E. (1987), Standards for validating health measures: Definition and content, Journal of Chronic Diseases, 40, 6, 473-480.

Ware, J.E. (1995), "The Status of Health Assessment 1994", Annual Review of Public Health, 16, 327-354.

Ware, J.E., R.H. Brook, A. Davies-Avery, K.N. Williams, A.L. Stewart, W.H. Rogers, C.A. Donald and S.A. Johnston (1980), "Conceptualization and measurement of health for adults in the Health Insurance Study. Volume I: Model of health and methodology", The Rand Corporation, R-1987/1-HEW, Santa Monica CA, USA.

Ware, J.E. and B. Gandek (1998), "Overview of the SF-36 health survey and the international quality of life assessment (IQOLA) project", Journal of Clinical Epidemiology, 51, 903-912.

Ware, J.E. and C.D. Sherbourne (1992), "The MOS 36 item short form health survey (SF-36)", Medical Care, 30, 473-483A.

Ware, J.E., K.K. Snow and M. Kosinski (1993, 2000), "SF-36 Health Survey: Manual and Interpretation Guide", QualityMetric Incorporated, Lincoln RI, USA.

Ware, J.E., M. Kosinski and S.D. Keller (1994), "SF-36 Physical & Mental Health Summary Scales: A User's Manual", Health Assessment Lab, New England Center, Boston MA, USA.

Ware, J.E., M. Kosinski, M.S. Bayliss, C.A. McHorney, W.H. Rogers and A. Raczek (1995), "Comparisons of Methods for the Scoring and Statistical Analysis of SF-36 Health Profile and Summary Measures: Summary of Results from the Medical Outcomes Study", Medical Care, 33, 4, AS264-AS279.

WHO (1999), "Anatomical Therapeutic Chemical (ATC) classification index", WHO Collaborating Centre for Drug Statistics Methodology, Oslo, Norway.

Wooldridge, J.M. (2001), Econometric analysis of cross section and panel data", 1st edition, The MIT Press, Cambridge MA, USA.

Sv.

SAMENVATTING

In competitieve markten voor individuele zorgverzekeringen laten risicoafhankelijke premies verschillen zien tussen subgroepen van verzekerden die worden bepaald door risicofactoren zoals leeftijd, geslacht, gezinsomvang, geografisch gebied, beroep, lengte van de contracttermijn, individuele of collectieve contracten, de hoogte van het vrijwillige eigen risico, gezondheidsstatus op het moment van inschrijving, leefstijl (roken, drinken, bewegen) en – via gedifferentieerde bonussen voor meerjarige no-claim – kosten in het verleden (Van de Ven et al. 2000). Financiële overdrachten zijn nodig om te voorkomen dat degenen met een hoog gezondheidsrisico in financiële zin moeite krijgen om zich te verzekeren. De beste oplossing om de betaalbaarheid van verzekeringsdekking voor hogerisicoverzekerden te verbeteren is een sponsor te vinden die een gereguleerd systeem van risicoafhankelijke premiesubsidies organiseert (Van de Ven et al. 2000). De financiële overdrachten lopen dan via een zogenaamd risicovereveningsfonds (REF). Door premiesubsidies zal de prijsconcurrentie tussen verzekeraars niet worden verstoord en daarom blijven de prikkels tot doelmatigheid onverminderd bestaan. In alle landen waar risicoafhankelijke premiesubsidies in hun zorgverzekeringsmarkten voorkomen worden deze in de vorm van risicoverevening tussen verzekeraars georganiseerd.

Hoewel de hoogte van de premies naar allerlei subgroepen kan worden gedifferentieerd, zal een sponsor niet alle premievariatie willen subsidiëren die in de praktijk voorkomt. De totale verzameling van risicofactoren die verzekeraars gebruiken om de hoogte van hun premies te differentiëren kan in twee deelverzamelingen worden gesplitst: de deelverzameling van risicofactoren die tot premievariatie leiden die de sponsor besluit te subsidiëren, de S(ubsidie)-type risicofactoren; en de deelverzameling van risicofactoren die tot premievariatie leiden die de sponsor besluit niet te subsidiëren, de N(on-subsidie)-type risicofactoren (Van de Ven en Ellis 2000, p. 768-769). In de meeste landen zullen waarschijnlijk geslacht, gezondheidsstatus en leeftijd tot op zekere hoogte als S-type risicofactoren worden beschouwd. Voorbeelden van mogelijke N-type risicofactoren zijn een grote geneigdheid tot medische consumptie, in een regio wonen met hoge prijzen en/of overcapaciteit wat resulteert in aanbodgeïnduceerde vraag of het gebruik van zorgaanbieders met een ondoelmatige wijze van praktijkvoering (Van de Ven et al. 2000). De sponsor bepaalt de specifieke indeling in S-type en N-type risicofactoren. In het geval dat de overheid de rol van de sponsor op zich neemt, wordt deze indeling uiteindelijk bepaald door de waardeoordelen van de samenleving.

Volgens de geldende wetgeving heeft de Nederlandse overheid gekozen om alleen kruissubsidies te geven voor kostenvariatie tussen subgroepen die zijn gebaseerd op de S-type risicofactoren leeftijd, geslacht en gezondheidsstatus (MoHWS 2005, p. 23). Leeftijd en geslacht zijn beschikbaar in de administraties van alle Nederlandse

verzekeraars en kunnen daarom relatief eenvoudig in de Nederlandse REF vergelijking (2.1) worden ingevoegd. Echter, de empirische mogelijkheden om een REF criterium voor gezondheid op individueel verzekerdenniveau in de REF vergelijking op te nemen zijn nogal beperkt. De Nederlandse REF vergelijking bevat in 2004 al een indrukwekkende reeks van op gezondheid gebaseerde administratieve criteria, inclusief de Farmacie Kosten Groepen (FKGs) en de Diagnose Kosten Groepen (DKGs). Hedentendage is dit het meest uitgebreide REF model op individueel verzekerdenniveau in de wereld. Het is echter nog steeds een open vraag in hoeverre zelfs deze uitgebreide verzameling van REF criteria risicoafhankelijke premiesubsidies (of: kruissubsidies) genereert die overeenkomen met de beleidsdoelen van de Nederlandse overheid. De centrale vraag van deze studie is daarom:

"In hoeverre leidt het Nederlandse risicovereveningsmodel uit 2004 tot risicoafhankelijke premiesubsidies die met de door de Nederlandse overheid geformuleerde beleidsdoelen overeenkomen en (hoe) kunnen deze subsidies worden verbeterd?"

Om een antwoord op deze centrale vraag te geven, worden drie onderzoeksvragen geformuleerd:

1. Gegeven de definitie van het basispakket, hoe kunnen we de kruissubsidies berekenen zoals die door de Nederlandse overheid zijn bedoeld? (Hoofdstukken 3, 4 en 5)
2. In hoeverre kunnen de bedoelde kruissubsidies worden benaderd door de operationalisaties voor gezondheidsstatus die in 2004 in de Nederlandse REF vergelijking zijn opgenomen? (Hoofdstuk 6)
3. In hoeverre kunnen de bedoelde kruissubsidies worden benaderd door alternatieve specificaties van het Nederlandse REF model uit 2004 of door premieregulering? (Hoofdstukken 7 en 8)

De belangrijkste bijdrage van deze studie is de ontwikkeling en empirische toepassing van een theoretisch raamwerk om de mate te bepalen waarin REF modellen tot de beoogde kruissubsidies leiden. Er wordt hierbij uitgegaan van periodieke acceptatieplicht van verzekerden door verzekeraars voor het gespecificeerde basispakket van vergoedingen en het bestaan van een systeem van risicoafhankelijke premiesubsidies. Echter, in deze studie wordt aangenomen dat de hoogte van de premies in de competitieve markt voor individuele zorgverzekeringen niet is gereguleerd.

In deze studie wordt een procedure ontwikkeld om te toetsen of een gegeven verzameling van REF criteria adequate compensatie biedt voor kostenvariatie

die door S-type risicofactoren wordt veroorzaakt (Hoofdstuk 1). Er wordt een overzicht gegeven van de relevante literatuur op het gebied van (hoofdzakelijk) administratieve operationalisaties van gezondheidsstatus die thans in gebruik of in onderzoek zijn. De in latere hoofdstukken toe te passen methodologie wordt nader uitgewerkt en in wiskundige termen beschreven. Verder worden richtlijnen gegeven voor de interpretatie van de resultaten en de vergelijkingen van de resultaten die het meest relevant bij het vinden van een antwoord op de centrale vraag van deze studie (Hoofdstuk 2).

Om de voorgestelde toetsprocedure toe te kunnen passen is een op maat gesneden gezondheidsenquête onder meer dan 50.000 ziekenfondsverzekerden uitgezet, zodanig dat hun gezondheidsprofiel veel preciezer kan worden beschreven dan als alleen van de REF criteria gebruik gemaakt kan worden (Hoofdstuk 3). De gezondheidsstatus variabelen zijn in deze studie uitgebreid getoetst op compleetheid, betrouwbaarheid en validiteit (Hoofdstuk 4).

In deze studie wordt aangenomen dat de Nederlandse overheid uitsluitend kruissubsidies wil voor geobserveerde kostenvariatie die wordt veroorzaakt door de S-type risicofactoren leeftijd, geslacht en gezondheidsstatus. Met betrekking tot een beperkte steekproef van gerespondeerde verzekerden wordt een alternatief risicovereveningsmodel op individueel verzekerdenniveau opgezet dat deze kostenvariatie zo goed als mogelijk beschrijft door alle operationalisaties van de S-type risicofactoren te gebruiken die in de gezondheidsenquête en administratieve bronnen aanwezig zijn (Hoofdstuk 5). De zogenaamde normatieve kosten die volgen uit dit alternatieve risicovereveningsmodel worden dan vergeleken met de REF voorspelde kosten die op de verzameling van REF criteria zijn gebaseerd die zijn meegenomen in de specificatie van de Nederlandse REF vergelijking uit 2004: leeftijd, geslacht, verzekeringsgrond, regio, FKGs en DKGs (Hoofdstuk 6). Aan de hand van dezelfde toetsprocedure kunnen tevens de juistheid van de kruissubsidies uit alternatieve specificaties van het REF model worden getoetst (Hoofdstuk 7). Ten slotte wordt aangetoond dat een verbetering van het REF model als strategie de voorkeur verdient om de financiële toegang tot verzekeringsdekking te vergroten dan impliciete kruissubsidies die door premierestricties worden gecreëerd (Hoofdstuk 8).

Een antwoord op de eerste onderzoeksvraag

De eerste onderzoeksvraag is hoe de kruissubsidies kunnen worden berekend zoals deze door de Nederlandse overheid zijn bedoeld, uitgaande van de bestaande invulling van het basispakket. In Hoofdstuk 5 worden de normatieve kosten voor een beperkte groep verzekerden (N=18.617) bepaald onder de veronderstelling dat de samenleving kruissubsidies wil voor kostenverschillen die door de S-type risicofactoren leeftijd, geslacht en gezondheidstoestand worden bepaald. De

normatieve kosten volgen uit een lineaire regressie van waargenomen kosten op een uitgebreide reeks van gezondheidsvariabelen uit de gezondheidsenquête en administratieve bronnen onder de aanname dat deze een adequate weergave van de S-type risicofactoren vormen.

Uit het literatuuroverzicht in Hoofdstuk 2 blijkt dat de meest veelbelovende kandidaten bij de operationalisatie van gezondheidstoestand in de REF vergelijking zelfgerapporteerde metingen van ervaren gezondheid, functionele gezondheid en chronische aandoeningen zijn. Als leidraad bij de specifieke keuze uit de operationalisaties van gezondheidsstatus wordt het conceptuele model van Ruwaard en Kramers (1997) toegepast. De gekozen operationalisaties van gezondheidsstatus zijn de acht SF-36 schalen fysieke gezondheid (PF), rol-fysieke gezondheid (RP), lichamelijke pijn (BP), algemene gezondheid (GH), vitaliteit (VT), sociaal functioneren (SF), rol-emotioneel (RE) en mentale gezondheid (MH), drie categorieën gebaseerd op het aantal OECD beperkingen (horen, zien en bewegen) en drie categorieën gebaseerd op het aantal specifieke zelfgerapporteerde chronische aandoeningen. De SF-36 is een 36-item instrument voor het meten van gezondheidstoestand en uitkomsten vanuit het patiëntperspectief en is ontwikkeld voor gebruik in de klinische praktijk en onderzoek, evaluaties van gezondheidsbeleid en enquêtes onder algemene populaties (Ware en Hayes 1988, Aaronson et al. 1998). FKGs en DKGs zijn ten slotte aan de normatieve kostenvergelijking toegevoegd omdat het niet noodzakelijkerwijs zo is dat hogere kosten van medische zorg samengaan met lagere scores op de hierboven genoemde gezondheidsschalen (Newhouse 1989).

In Hoofdstuk 3 worden de data beschreven die in deze studie zijn gebruikt. Een uitgebreide reeks van gezondheidsvariabelen is door middel van een op maat gesneden gezondheidsenquête verkregen die in 2001 onder 50.022 Agis verzekerden is gehouden. De bruto respons op de zgn. Agis Gezondheidsenquête 2001 was 23.163 (46,3%). Voor het doel van dit onderzoek zijn 18.617 records in het onderzoeksbestand geschikt omdat voor deze respondenten valide SF-36 scores konden worden afgeleid en de administratieve gegevens uit de jaren 2001 en 2002 zowel beschikbaar als valide waren. Validiteit en betrouwbaarheid van de acht SF-36 schalen wordt getoetst en akkoord bevonden in Hoofdstuk 4. Verder is een gegevensverzameling met een panelstructuur voor de jaren 1997-2002 afgeleid uit de ziekenfondsadministratie van Agis Zorgverzekeringen (en haar voorgangers) en zijn additionele onderzoeksgegevens beschikbaar gesteld door Nederlandse onderzoeksbedrijven (APE Public Economics en Prismant, respectievelijk gevestigd te Den Haag en Utrecht).

Normatieve kosten volgen uit de met lineaire regressie geschatte vergelijking (2.4) als zijnde de kosten die in 2002 worden verwacht uitgaande van de subgroepen

die zijn gebaseerd op de zogenaamde S-type criteria leeftijd en geslacht, de niet gerangordende versies van de FKGs en de DKGs, de acht SF-36 schalen, het aantal OECD beperkingen en het aantal zelfgerapporteerde chronische aandoeningen. Merk op dat, in tegenstelling tot de (gerangordende versies van de) FKGs en DKGs in de REF vergelijking van 2004, verzekerden in de normatieve vergelijking in meerdere FKGs en DKGs tegelijkertijd kunnen worden ingedeeld. Een antwoord op de eerste onderzoeksvraag is dat kruissubsidies zoals bedoeld door de sponsor voor een beperkte steekproef van respondenten op de enquête kunnen worden berekend door de aanpak te volgen die in deze studie is ontwikkeld.

Een antwoord op de tweede onderzoeksvraag

Om de tweede onderzoeksvraag te beantwoorden moet worden nagegaan in welke mate de REF criteria in de Nederlandse REF vergelijking van 2004 leiden tot de kruissubsidies zoals de Nederlandse overheid ze heeft bedoeld. Om exact in overeenstemming te zijn met de beleidsdoelen van de sponsor worden de kruissubsidies idealiter gebaseerd op de normatieve kosten. In dat hypothetische geval geldt voor de subgroepen die door de S-type criteria worden gevormd dat de REF verwachte kosten gelijk zijn aan de normatieve kosten. Vandaar dat de daadwerkelijk in de praktijk ingevoerde kruissubsidies kunnen worden getoetst door na te gaan in hoeverre de REF voorspelde kosten afwijken van deze normatieve kosten voor de subgroepen zoals gedefinieerd door de S-type criteria.

In Hoofdstuk 6 wordt een toets uitgevoerd op de criteria die in 2004 in de Nederlandse REF vergelijking zijn opgenomen, uitgaande van de normatieve kosten zoals afgeleid in Hoofdstuk 5. Uit vergelijking (2.6) volgt dat de gewogen gemiddelde afwijkingen van de REF voorspelde kosten ten opzichte van de normatieve kosten voor de subgroepen gedefinieerd door de S-type criteria gelijk is aan 198. Dit getal zou 687 zijn geweest indien er helemaal geen criteria in vergelijking (2.1) zouden zijn opgenomen, met andere woorden indien de REF verwachte kosten voor alle verzekerden gelijk zijn aan de gemiddelde waargenomen kosten in de totale onderzoekspopulatie. Hieruit volgt dat 1 – (198/687) x 100% = 71.2% van de door de Nederlandse overheid beoogde kruissubsidies kunnen worden bereikt met de REF criteria die in 2004 in de Nederlandse REF vergelijking voorkomen. Deze vaststelling vormt het antwoord op de tweede onderzoeksvraag.

REF voorspelde kosten en normatieve kosten kunnen ook worden vergeleken voor subgroepen die door de REF criteria worden bepaald (in plaats van de S-type criteria). Afwijkingen van de REF voorspelde kosten ten opzichte van de normatieve kosten mogen worden toegerekend aan N-type risicofactoren. REF voorspelde kosten voor arbeidsongeschikte verzekerden blijken met 420 euro (15,1%) te

worden overgecompenseerd, bijstandsgerechtigde verzekerden en zelfstandige ondernemers worden ondergecompenseerd met respectievelijk 327 euro (16,2%) en 162 euro (16,2%) ten opzichte van normatieve kosten. Verzekerden woonachtig in het eerste regionale cluster van ZIP codes worden gemiddeld met 245 euro (13,6%) overgecompenseerd ten opzichte van normatieve kosten, terwijl verzekerden die in de regionale clusters 6, 7 en 8 wonen respectievelijk met 190 euro (10,7%), 94 euro (5,4%) en 171 euro (11,4%) worden ondergecompenseerd. Ten aanzien van de 2004 specificatie van de Nederlandse REF vergelijking moet de aanname daarom worden verworpen dat deze REF criteria "zuivere" operationalisaties van gezondheidsverschillen zijn, d.w.z. deze REF criteria leiden in de praktijk onbedoeld tot compensatie voor N-type kostenvariatie. Het is in deze context belangrijk op te merken dat de onderzoeksresultaten in deze studie betrekking hebben op verzekerden van slechts één verzekeraar, die daarom niet kunnen worden beschouwd als representatief voor alle Nederlandse provincies.

Verder blijkt dat REF voorspelde kosten voor zelfstandige ondernemers lager uitvallen dan die voor werknemers in loondienst, terwijl dat niet geldt voor de normatieve kosten van deze subgroepen. Blijkbaar leidt toepassing van het REF criterium verzekeringsgrond in onvoldoende mate tot compensatie voor S-type kostenvariatie. Er zij opgemerkt dat dit probleem niet kan worden opgelost door de subgroepen van werknemers in loondienst en van zelfstandige ondernemers tot één subgroep te combineren, een oplossing waarover hevig gedebatteerd werd in de context van de 2004 specificatie van de Nederlandse REF vergelijking.[153] Uiteindelijk besloot de Nederlandse overheid om werknemers in loondienst en zelfstandige ondernemers als aparte subgroepen in de REF vergelijking van 2004 op te nemen, onder de veronderstelling dat het verschil hoofdzakelijk aan de S-type risicofactor gezondheid kon worden toegeschreven. Op basis van de resultaten in deze studie blijkt nu echter dat dit besluit om de lagere risicoafhankelijke premiesubsidies voor zelfstandige ondernemers te genereren ingaat tegen de beleidsdoelen zoals de Nederlandse overheid die heeft geformuleerd.

Idealiter wordt een afwijking tussen REF voorspelde kosten en de normatieve kosten zoals bij zelfstandige ondernemers weggenomen zodanig dat de REF criteria niet langer kruissubsidies voor N-type kostenvariatie genereren. Dit kan worden bereikt door een aanpassing van de REF gewichten, bijvoorbeeld door toepassing van de zogenaamde weggelaten variabelen benadering dan wel de zogenoemd normatieve aanpassingsprocedure die in deze studie is ontwikkeld. De weggelaten variabelen benadering om N-type vertekening van de REF gewichten

153. Een dergelijke oplossing zou namelijk ten koste gaan van lagere risicoafhankelijke premiesubsidies voor de werknemers in loondienst.

te verwijderen is voorgesteld door Schokkaert, Dhaene en Van de Voorde (1998) en Schokkaert en Van de Voorde (2004). Echter, de afname van het verschil tussen REF voorspelde kosten en normatieve kosten blijkt nogal beperkt bij deze methode, tenminste gegeven het onderzoeksbestand dat in deze studie is gebruikt en gegeven de specifieke operationalisatie van N-type criteria. Een alternatieve methode om de REF gewichten aan te passen is om deze te vervangen door de zogenoemd normatief aangepaste REF gewichten uit vergelijking (2.8). In dat geval wordt alle N-type vertekening van de onaangepaste REF gewichten weggenomen.

Aanpassing van de REF gewichten volgens de weggelaten variabelen benadering leidt tot risicoafhankelijke premiesubsidies die voor (1-201/687) x 100% = 70.7% aan de door de Nederlandse overheid gestelde doelen beantwoorden. De normatieve aanpassing van de REF gewichten geeft een uitkomst van (1-209/687) x 100% = 69.6%. Blijkbaar gaat het verwijderen van N-type vertekening uit de REF gewichten gepaard met een beperkte afname van de mate waarin de REF criteria tot compensatie voor S-type kostenvariatie leiden. Daarom wordt aanbevolen om normatief aangepaste gewichten in een REF vergelijking te gebruiken in plaats van onaangepaste REF gewichten indien de afname van compensatie voor S-type kostenvariatie niet al te groot is.

Een antwoord op de derde onderzoeksvraag

Om de derde onderzoeksvraag te beantwoorden moet worden bepaald in welke mate de beoogde kruissubsidies kunnen worden verkregen door toepassing van alternatieve specificaties van het Nederlandse REF model uit 2004 of door premieregulering. In Hoofdstuk 6 zijn de REF gewichten aangepast om zo het verschil tussen REF voorspelde kosten en normatieve kosten te verkleinen voor de subgroepen van de REF criteria. Echter, bij deze aanpassing van gewichten is de verzameling van toegepaste REF criteria uit de REF vergelijking van 2004 ongewijzigd gebleven. In Hoofdstuk 7 worden alternatieve specificaties van het Nederlandse REF model uit 2004 getoetst om het verschil te reduceren tussen REF voorspelde kosten en normatieve kosten voor de subgroepen zoals bepaald door de S-type criteria. Deze procedure kan leiden tot kruissubsidies die beter aansluiten op de beleidsdoelen van de sponsor.

Als eerste voorbeeld worden zogenoemde paramedische kosten groepen (PMKGs), medische hulpmiddelen kosten groepen (MHKGs) en mentale farmacie kostengroepen (MFKGs) geconstrueerd op basis van administratieve data. Paramedische indicatoren van chronische aandoeningen worden gebruikt als indicatoren van fysieke beperkingen, medische hulpmiddelen als indicatoren van functionele problemen en geneesmiddelen die op het zenuwstelsel inwerken als indicatoren

van geestelijke aandoeningen. Deze mogelijk nieuwe REF criteria zijn afgeleid van de declaratiegegevens uit de ziekenfondsadministratie van Agis. Toevoeging van de PMKGs, MHKGs en MFKGs aan de specificatie van de Nederlandse REF vergelijking uit 2004 blijkt tot een substantiële reductie van het gat tussen REF voorspelde kosten en normatieve kosten te leiden. Op basis van vergelijking (2.6) blijkt het gewogen gemiddelde van de absolute verschillen tussen REF voorspelde kosten en normatieve kosten voor de subgroepen van de S-type criteria gelijk te zijn aan 157. Gegeven dat dit 687 zou zijn bij afwezigheid van een systeem van risicoafhankelijke premiesubsidies betekent dit dat (1 – 157/687) x 100% = 77.1% van de bedoelde kruissubsidies kan worden behaald door toevoeging van de bovengenoemde REF criteria aan de Nederlandse REF vergelijking van 2004. Merk op dat dit getal gelijk is aan 71.2% zonder toevoeging van deze nieuwe REF criteria. Er kan worden geconstateerd dat toevoeging van PMKGs, MHKGs en MFKGs tot een substantiële verbetering van de risicoafhankelijke premiesubsidies zal leiden, alhoewel er ruimte voor verbetering blijft bestaan.

Als tweede voorbeeld wordt een specifieke vorm van ex-post risicodeling getoetst als supplement voor incomplete en/of imperfecte REF criteria. Er wordt een vorm toegepast dat in beperkte mate een variant is van het systeem van ex-post risicodeling zoals dat in 2004 in Nederland werd toegepast: 100% retrospectieve vergoeding van de vaste kosten ziekenhuisverpleging en 90% compensatie van de kosten variabele kosten ziekenhuisverpleging, specialistische hulp en overige prestatie boven een drempel van EURO 12.500. In deze studie blijkt het gat tussen REF voorspelde kosten en normatieve kosten hierdoor voor een groot deel te kunnen worden gereduceerd. Het gewogen gemiddelde van de verschillen tussen REF voorspelde kosten en normatieve kosten voor de subgroepen die zijn gebaseerd op de S-type criteria is in dit geval gelijk aan 141. Daarom kan (1 – 141/687) x 100% = 79.5% van de beoogde kruissubsidies worden behaald door toepassing van het hiervoor genoemde systeem van ex-post risicodeling als supplement bij de Nederlandse REF vergelijking in 2004. Ex-post risicodeling zal ook een belangrijke bijdrage blijven leveren na toevoeging van de PMKGs, MHKGs en MFKGs aan de REF vergelijking. Er zij opgemerkt dat de invoering van een ex-post arrangement in de praktijk zal leiden tot een vermindering van de prikkels tot doelmatigheid en daarom tot een te maken afweging met de gewenste verbetering van de premiesubsidies.

Als derde voorbeeld worden de REF gewichten afgeleid binnen een het kader van een GLM model onder de veronderstelling van een Gamma verdeling en een log link tussen REF voorspelde kosten en de REF criteria. De afstand tussen REF voorspelde kosten en normatieve kosten blijkt aanmerkelijk te worden verkleind voor bijna alle subgroepen die door de S-type criteria worden gedefinieerd, hoewel

deze afname kleiner is dan bij de hierboven genoemde varianten van het Nederlandse REF model uit 2004. Het gewogen gemiddelde van de afwijkingen van REF voorspelde kosten van de normatieve kosten is 172 voor de subgroepen die zijn gebaseerd op de S-type criteria. Daarmee wordt (1 – 172/687) x 100% = 75.0% van de beoogde kruissubsidies behaald als gevolg van deze aangepaste statistische specificatie van het REF model.

Door deze hele studie heen is aangenomen dat er sprake is van periodieke acceptatieplicht voor verzekeraars ten aanzien van het basispakket van vergoedingen, zijn verzekeraars vrij om de hoogte van hun premies te differentiëren en is er sprake van een systeem van risicoafhankelijke premiesubsidies dat door een sponsor wordt georganiseerd om de financiële toegang tot verzekeringsdekking voor hogerisicoverzekerden te garanderen. Echter, geen van de specificaties van het REF model blijkt tot kruissubsidies te leiden die volledig compenseren voor S-type kostenvariatie. Daarom kan de sponsor in de verleiding komen om premierestricties in te voeren om alsnog de financiële toegang tot verzekeringsdekking voor hogerisicoverzekerden veilig te stellen.

Premies zijn de directe betalingen die verzekerden doen aan hun verzekeraars en de premiestelling kan worden gereguleerd door de sponsor. Premieregulering kent verschillende vormen: een uniforme premie per verzekeraar, een verbod op het gebruik van bepaalde risicofactoren bij de premiestelling of het instellen van een bandbreedte (per premieklasse). Het alleen toestaan van uniforme premies per verzekeraar is de meest extreme vorm van premieregulering, omdat iedereen die bij eenzelfde verzekeraar is verzekerd dezelfde premie moeten betalen voor het basispakket onafhankelijk van het individuele risicoprofiel. Echter, hoewel premierestricties zijn bedoeld om impliciete kruissubsidies te genereren voor S-type kostenvariatie, leiden zij ook tot impliciete kruissubsidies voor kostenvariatie veroorzaakt door N-type risicofactoren hetgeen per definitie in conflict is met de doelen die de sponsor beoogt. Gegeven het in deze studie ontwikkelde theoretische raamwerk, is het mogelijk om precies te bepalen hoeveel impliciete kruissubsidie er als gevolg van premierestricties tussen subgroepen voor S-type kostenvariatie wordt gegenereerd en hoeveel voor N-type kostenvariatie.

Hoofdstuk 8 laat zien dat een uniforme premie voor de meeste subgroepen die zijn gebaseerd op zelfgerapporteerde historische medische consumptie, gezondheidsstatus, ziektes en aandoeningen tot kruissubsidies voor voornamelijk S-type kostenvariatie leidt. Echter, de voorspelbare winsten en verliezen blijken voornamelijk bepaald door N-type risicofactoren voor de subgroepen van verzekerden die zijn samengesteld op basis van het aantal jaren dat zij ieder jaar voorafgaand aan 2002 tot de 25% verzekerden met de meeste zorgkosten behoorden. Hieruit volgt

dat voor sommige subgroepen van verzekerden de premierestricties impliciete kruissubsidies genereren die grotendeels in overeenstemming zijn met de beleidsdoelen van de sponsor, maar voor andere subgroepen zijn deze kruissubsidies juist overwegend strijdig met die beleidsdoelen.

Nederlandse verzekeraars is het volgens de Zorgverzekeringswet 2006 toegestaan om de hoogte van hun premies te differentiëren naar de twaalf provincies. Uit de resultaten van Hoofdstuk 8 blijkt nu dat, gegeven de specificatie van de Nederlandse REF vergelijking uit 2004, de mogelijke variatie in de hoogte van de provinciale premies hoofdzakelijk kan worden toegerekend aan N-type risicofactoren. Dit blijkt zelfs nog meer het geval te zijn indien de vertekening door N-type risicofactoren uit de REF gewichten is verwijderd. Dit onderzoeksresultaat rechtvaardigt het besluit van de Nederlandse overheid om premieverschillen naar de twaalf provincies toe te staan: compensatie voor regionale S-type kostenvariatie blijkt al adequaat door de REF vergelijking te worden gegenereerd. Hierbij dient opgemerkt te worden dat de onderzoeksgegevens in deze studie betrekking hebben op de verzekerden van slechts één verzekeraar, die daarom niet kunnen worden beschouwd als representatief voor alle Nederlandse provincies.

Het antwoord op de derde onderzoeksvraag van deze studie is dat alternatieve specificaties van het REF model uit 2004 tot een substantiële verbetering van de kruissubsidies kunnen leiden. Tot 79.5% van de door de Nederlandse overheid beoogde kruissubsidies kan worden bereikt door toepassing van een systeem van ex-post risicodeling, echter een nadeel van een dergelijk systeem is dat dit ten koste gaat van de prikkel tot doelmatigheid. Dit nadeel bestaat niet in geval van toevoeging van nieuwe criteria zoals PMKGs, MHKGs en MFKGs aan de verzameling van REF criteria uit 2004 of als een multiplicatieve in plaats van additieve specificatie van de REF vergelijking wordt toegepast. In deze laatste gevallen kunnen respectievelijk 77.1% en 75.0% van de beoogde kruissubsidies worden bereikt.

Als de verbeterde kruissubsidies nog niet volledig beantwoorden aan de beleidsdoelen van de sponsor, dan kunnen premierestricties zorgen voor de benodigde kruissubsidiëring tussen de subgroepen die op de S-type risicofactoren zijn gebaseerd. Echter, voor andere subgroepen van verzekerden kunnen deze premierestricties ook leiden tot niet beoogde kruissubsidies voor N-type kostenvariatie. Tegelijkertijd ontstaan ook prikkels tot selectie met mogelijk negatieve effecten op kwaliteit van de zorg, betaalbaarheid van de verzekeringsdekking voor hogerisicoverzekerden en doelmatigheid in de productie van zorg (Van de Ven, Van Vliet en Lamers 2004). Toepassing van premierestricties als aanvulling op de risicoafhankelijke premiesubsidies moet daarom noodzakelijk en proportioneel

zijn voor het behalen van de gestelde beleidsdoelen en vereist een zorgvuldige afweging tegen de prikkels tot selectie (en hun mogelijk negatieve gevolgen) die hiermee worden geïntroduceerd.

Een antwoord op de centrale vraag

Het antwoord op de centrale vraag van deze studie is dat de REF criteria in de Nederlandse REF vergelijking van 2004 tot kruissubsidies leiden die voor 71.2% overeenkomen met de beleidsdoelen die de Nederlandse overheid zich heeft gesteld. Dit resultaat kan worden verbeterd tot 79.5% door ex-post risicodeling, hetgeen moet worden afgewogen tegen de mate waarin sprake is van een vermindering van de prikkel tot doelmatigheid. Een alternatieve strategie om de kruissubsidies van de REF vergelijking uit 2004 te verbeteren kan bestaan uit het toevoegen van nieuwe criteria zoals PMKGs, MHKGs en MFKGs, in welk geval de prikkel tot doelmatigheid onveranderd blijft. Verder wordt aanbevolen aanvullend onderzoek te doen naar het hanteren van een multiplicatieve in plaats van additieve specificatie van de REF vergelijking. Bij voorkeur worden premierestricties alleen toegepast voorzover ze noodzakelijk en proportioneel zijn als aanvulling op de risicoafhankelijke premiesubsidies om de specifieke beleidsdoelen van de sponsor te kunnen behalen en dient een expliciete afweging plaats te vinden ten aanzien van de prikkels tot selectie die door premierestricties ontstaan. Het besluit van de Nederlandse overheid om sinds 2006 premiedifferentiatie naar de twaalf provincies toe te staan, blijkt in dit opzicht goed te rechtvaardigen. In het algemeen wordt, gegeven een specifieke keuze van de REF criteria, aanbevolen om in een REF vergelijking de normatief aangepaste REF gewichten te gebruiken in plaats van de onaangepaste REF gewichten, hoewel steeds in de gaten moet worden gehouden dat de compensatie van de kruissubsidies voor S-type kostenvariatie daarbij zoveel mogelijk gehandhaafd blijft.

Algemene beleidsaanbevelingen

De modelspecificaties in de Hoofdstukken 6, 7 en 8 worden gebruikt om de toepassing te illustreren van de toetsprocedure die in deze studie is ontwikkeld. Het theoretische raamwerk dat in deze studie is ontwikkeld, kan worden toegepast op alle specificaties van risicovereveningsmodellen in competitieve markten voor individuele zorgverzekeringen om te beoordelen of de kruissubsidies die een sponsor in de praktijk heeft ingevoerd overeenstemmen met haar beleidsdoelen. In het algemeen kan dit theoretische raamwerk ook worden gebruikt in andere sectoren waarbij indirecte standaardisatie van populaties wordt bepaald door een normatieve beslisregel in plaats van louter een statistische exercitie.

De eerste beleidsaanbeveling is dat een sponsor bij invoering van een risicovereveningsmodel altijd een expliciete keuze dient te maken ten aanzien van de indeling in S-type en N-type risicofactoren. Een dergelijke expliciete keuze vormt het uitgangspunt voor de toepassing van het in deze studie ontwikkelde theoretische raamwerk om na te kunnen gaan in hoeverre een REF model de betaalbaarheid van verzekeringsdekking voor de hogerisicoverzekerden kan veilig stellen.

De tweede beleidsaanbeveling is het op reguliere basis toepassen van het theoretisch raamwerk dat in deze studie is ontwikkeld op een landelijke steekproef van verzekerden om de juistheid van de kruissubsidies te toetsen en na te gaan of de aangepaste REF gewichten in plaats van de onaangepaste REF gewichten moeten worden toegepast als de REF criteria in de praktijk imperfect zijn.

De derde beleidsaanbeveling is om de kruissubsidies van verzekerden met functionele beperkingen te verbeteren door de ontwikkeling van REF criteria die zijn gebaseerd op het gebruik van fysiotherapie, medische hulpmiddelen en medicijnen ten behoeve van geestelijke aandoeningen. In het bijzonder kan de laatste categorie REF criteria waardevol blijken als, in tegenstelling tot de kostendefinitie die in deze studie is gehanteerd, geestelijke gezondheidszorg ook in het Nederlandse basispakket wordt opgenomen.

De vierde beleidsaanbeveling is om bij incompleetheid van de verzameling REF criteria ex-post risicodeling toe te voegen als aanvulling op de REF vergelijking. Uit deze studie blijkt dat ex-post risicodeling de financiële toegang van hogerisicoverzekerden meer verhoogt dan het toevoegen van REF criteria gebaseerd op fysiotherapie, medische hulpmiddelen en medicijnen ten behoeve van geestelijke aandoeningen. Echter, ex-post risicodeling reduceert tegelijkertijd de prikkel tot doelmatigheid. Bij ex-post risicodeling is daarom sprake van een afweging tussen de juistheid van de kruissubsidies en de prikkels tot doelmatigheid.

De vijfde beleidsaanbeveling is om verzekeraars toe te staan dat zij de hoogte van hun premies mogen aanpassen aan bestaande N-type kostenvariatie. Premierestricties kunnen gelden ten aanzien van S-type risicofactoren voorzover de daarmee samenhangende kostenvariatie door de REF criteria en/of ex-post risicodeling wordt gesubsidieerd. Echter, de (impliciete) kruissubsidies die voortvloeien uit de premierestricties zijn per definitie niet bedoeld om te compenseren voor N-type risicofactoren, terwijl premierestricties wel leiden tot prikkels tot selectie met mogelijk negatieve effecten op kwaliteit, betaalbaarheid en doelmatigheid. De onterechte compensatie voor N-type kostenvariatie en (de negatieve effecten van) genoemde prikkels kunnen worden voorkomen als verzekeraars de hoogte van hun premies mogen aanpassen aan de kostenvariatie die door N-type risicofactoren wordt veroorzaakt.

Als de sponsor alleen kruissubsidies wil voor kostenvariatie die door S-type risicofactoren wordt veroorzaakt – bijvoorbeeld leeftijd, geslacht en gezondheid – dan zou een proportionele implementatie van premierestricties eruit kunnen bestaan dat premiedifferentiatie naar factoren die operationalisaties zijn van de S-type risicofactoren (inclusief de REF criteria) wordt verboden, maar niet naar premiefactoren die operationalisaties zijn van de N-type risicofactoren zoals dat wel het geval is bij doorsneepremies. Een minder beperkend alternatief voor doorsneepremies per verzekeraar per product kan zijn het instellen van bandbreedtes waarbinnen een verzekeraar de premiehoogte mag vaststellen. Gegeven de vrijheid die een bandbreedte biedt, zullen verzekeraars in dat geval signalen afgeven welke relevante operationalisaties van de S-type risicofactoren geschikt zijn om in de dooropvolgende jaren aan de REF vergelijking toe te voegen. De sponsor kan doorsneepremies per risicogroep of klasse invoeren om de consumenten tegen te sterke premiestijgingen te beschermen in geval van nieuwe premiefactoren. In ieder geval blijft het de vraag waarom de Nederlandse overheid verzekeraars nog steeds verbiedt hun premies te differentiëren op basis van de operationalisaties van N-type risicofactoren, zoals regionale input prijzen, wijze van praktijkvoering en consumptiegeneigdheid.

Land-specifieke beleidsaanbevelingen

De Nederlandse overheid is volgens de Zorgverzekeringswet 2006 verplicht om in 2008 en 2011 een wetenschappelijke evaluatie van het REF model uit te laten voeren door een panel van internationale experts (MoHWS 2005, pagina 26). Het in deze studie ontwikkelde theoretisch raamwerk kan voor dit doel worden gebruikt om de juistheid van de kruissubsidies te bepalen voor een gestratificeerde steekproef uit de totale Nederlandse bevolking. Een dergelijke steekproef kan worden gevonden in het Permanent Onderzoek LeefSituatie (POLS) dat jaarlijks wordt gehouden door het Centraal Bureau voor de Statistiek (CBS). De module 'Gezondheid' komt grotendeels overeen met de Agis Gezondheidsenquête 2001 en kan worden gebruikt om de normatieve kosten voor een nationale steekproef van verzekerden af te leiden (de SF-12 wordt door het CBS gebruikt als zijnde de verkorte versie van de SF-36).[154] Op basis van de verwachting dat in 2008 de geneeskundige geestelijke gezondheidszorg aan het basispakket zal worden

154. De reguliere frequentie waarmee deze toetsprocedure kan worden herhaald is niet vaker dan om de twee jaar, vanwege beperkingen in de steekproefomvang bij de landelijke enquête. Bij de module 'Gezondheid' in de jaarlijkse POLS enquête wordt gestreefd naar een netto respons van 10.000 respondenten, terwijl de Agis Gezondheidsenquête 18.617 records bevat. In het algemeen hangt de aanbevolen omvang van de steekproef in deze studie af van de mate van detail die nodig is om de REF vergelijking en de normatieve vergelijking op te stellen.

toegevoegd, mag worden verwacht dat het relatieve belang van de vier mentale SF-36 schalen in de normatieve vergelijking zal verschillen van dat in de voorliggende studie.

In Zwitserland worden leeftijd, geslacht en het kanton waarin de verzekerde woont als REF criteria gebruikt, zonder dat dit wordt aangevuld met een systeem van ex-post risicodeling (Beck et al. 2003, Van de Ven et al. 2007). In 1996 is wettelijk besloten dat de modelspecificatie voor een periode van 10 jaren onveranderd zou blijven, in 2004 is besloten om het REF model tot 2010 te handhaven (Bundesrat 2004, Bundesrat 2005). De Eerste Kamer heeft in 2006 met wetgeving ingestemd om ziekenhuis- en verpleeghuisopnamen uit het verleden (indien opgenomen voor ten minste drie dagen) als REF criterium aan de Zwitserse REF vergelijking toe te voegen (Ständerat 2006, Art. 18a, Lid 2, pagina 76). Verder gaat men akkoord met risicoverevening als permanente maatregel, nog nader te specificeren aanvullende gezondheidscriteria als lange termijn optie en de overstap van retrospectieve naar prospectieve berekening van de REF gewichten. In mei 2006 werd er na een hoorzitting in de Tweede Kamer nog geen besluit over deze wetgeving genomen, maar in 2007 wordt verwacht dat de Tweede Kamer ermee zal instemmen. Voor het eind van 2010 moet de Zwitserse overheid een besluit hebben genomen over invoering van gezondheidscriteria in de REF vergelijking (bijvoorbeeld FKGs), alsmede over de invoering van een systeem van ex-post risicodeling en permanente handhaving van het systeem van risicoverevening (BBI 2004 4259, pagina 4273). Het in deze studie ontwikkelde theoretisch raamwerk kan bij dit delicate Zwitserse debat van nut zijn, omdat het duidelijk maakt in hoeverre kruissubsidies noodzakelijk zijn om de problemen van het marktmechanisme te verminderen waarmee de Zwitserse overheid worden geconfronteerd (Van de Ven et al. 2007). Appendix A6.1 en Sectie 7.2 laten de mogelijke consequenties van deze besluiten zien op basis van de steekproef van Nederlandse verzekerden die in deze studie is gebruikt. Merk op dat de Zwitserse definitie van het basispakket ook verpleeghuiszorg en thuiszorg bevat. Het ligt in de lijn der verwachtingen dat de specificaties van het Nederlandse REF model zoals gepresenteerd in Hoofdstuk 6 en 7 minder geschikte kruissubsidies zouden genereren als verpleeghuiszorg en thuiszorg ook tot het Nederlandse basispakket zouden hebben behoord. Daarom mag worden verwacht dat de uitdagingen om in Zwitserland een geschikte modelspecificatie te vinden ten minste zo groot zullen zijn als die met betrekking tot het Nederlandse REF model.

In Duitsland was de invoering van morbiditeitgerelateerde REF criteria als operationalisatie van de S-type risicofactor gezondheidsstatus oorspronkelijk gepland in 2007, maar is nu ingepland voor 2009 en zal naar verwachting op 50-80 aandoeningen betrekking hebben (Büchner en Wasem 2003, Bundesrat 2007). Dan

zal ook worden overgeschakeld van een interne naar een externe wijze van de REF betalingen aan de ziekenfondsen worden overgeschakeld, d.w.z. naar modaliteit B zoals beschreven in Van de Ven et al. (2000, pagina 324). In 2002 is een systeem van ex-post risicodeling voor 60% van de kosten boven een drempel van meer dan 20.000 euro ingevoerd, waarmee het niveau van de ex-post kostencompensatie werd opgevoerd van 0% naar 4% (in 2006) (Van de Ven et al. 2007, Tabel 1). In 2003 zijn (de vrijwillige registraties voor) geaccrediteerde Disease-Management-Programma's (DMPs) als REF criterium aan de Duitse REF vergelijking toegevoegd. Zowel de ex-post risk sharing als de DMPs moeten als tijdelijke maatregelen worden gezien en zullen worden afgeschaft bij de geplande systeemwijziging in 2009. In 2004 adviseerde een internationale groep van experts om FKGs (RxGroups) en DKGs (IPHCC) voor dit doel te gebruiken (IGES/Lauterbach/Wasem 2004), in 2006 wordt er nog steeds een debat gevoerd over invoering van morbiditeitgerelateerde REF criteria (Schokkaert et al. 2006, Tabel 1). Het aparte verzekeringssysteem voor ziekenfondsverzekerden en particulier verzekerden zal na 2009 in stand worden gehouden, echter, het zal ziekenfondsen worden toegestaan om hogere premies bij hun verzekerden in rekening te brengen dan de landelijk verwachte premie waar de Duitse overheid vanuit zal gaan. De particuliere verzekeraars zullen te maken krijgen met een periodieke acceptatieplicht voor een basispakket analoog aan dat van de ziekenfondsen. Verder zal er een maximum aan de premiehoogte worden ingesteld die afhangt van de gemiddelde premiehoogte in de ziekenfondssector en mag er geen risico-opslag in de premie worden opgenomen. Vanwege deze premieregulering zal er enige vorm van risicopooling worden toegepast, maar de specifieke vorm blijft vooralsnog onbepaald. Het effect van invoering van FKGs en DKGs in de ziekenfondssector en het effect van (invoering van) een systeem van ex-post risicodeling in beide verzekeringssectoren kan met de in deze studie ontwikkelde toetsprocedure alvast worden bepaald.

In Israel is alleen leeftijd als REF criterium in de REF vergelijking opgenomen, sinds 2005 gaat het om elf in plaats van negen subgroepen. Geslacht kan (nog) niet gebruikt worden vanwege beschikbaarheidsproblemen en er is geen systeem van ex-post risicodeling (Shmueli, Chernichovsky en Zmora 2003). Er bestaat een groeiende ontevredenheid met deze formule: (1) er wordt betoogd dat kinderen worden overgecompenseerd en ouderen ondergecompenseerd, en (2) er wordt beargumenteerd dat er meer REF criteria moeten komen (Van de Ven et al. 2007). In 2006 gaat het debat over de invoering van morbiditeitgerelateerde REF criteria nog steeds voort (Schokkaert et al. 2006, Tabel 1). Arbeidsongeschiktheid is niet als REF criterium ingevoerd omdat hiervan wordt verondersteld dat dit geen relevant criterium is als operationalisatie van gezondheid. Verder dienen ook socio-economische kenmerken niet als REF criteria onder de veronderstelling dat

tijdprijs ("tijd om naar de dokter te gaan") een belangrijker bepalende factor van de waargenomen kosten is dan gezondheid. Echter, gebaseerd op de resultaten zoals gepresenteerd in Tabel 6.7 moet worden geconcludeerd dat er geen goede reden is om arbeidsongeschiktheid of specifieke socio-economische kenmerken in de Israëlische REF vergelijking op te nemen, gegeven dat een aanpassing van de REF gewichten kan worden toegepast om de N-type kostenvariatie zoals tijdprijs van verzekerden die zelfstandig ondernemer zijn. In 2005 is overeengekomen dat de Israëlische REF vergelijking na iedere vier jaren zou worden aangepast, daarom is in 2009 een mogelijke revisie van hun REF vergelijking mogelijk. In de tussentijd kan het theoretisch raamwerk dat in deze studie is ontwikkeld in Israël worden toegepast om tot onderzoeksresultaten te komen die op hun specifieke situatie betrekking hebben. Merk op dat de Israëlische formule niet op individuele claims data van de ziekenfondsen is gebaseerd, maar op een enquête met betrekking tot zorggebruik en een gegevensverzameling van ziekenhuisopnamen. Data met betrekking tot geneesmiddelen worden in het geheel niet betrokken in de berekeningen.

In België is de schatting van de REF gewichten sinds 2002 op individuele data gebaseerd. De Belgische REF criteria zijn leeftijd, geslacht, morbiditeitgerelateerde en socio-economische variabelen, bijvoorbeeld indicatoren van chronische ziekten en categorieën van arbeidsongeschiktheid. De informatie met betrekking tot DRGs en geneesmiddelengebruik is verzameld maar nog niet toegepast, maar er is al wel algemene consensus over de wenselijkheid om deze in het toekomstige REF model in te voegen (Schokkaert et al. 2006). In het voortdurende debat over de specifieke keuze van S-type en N-type risicofactoren wordt veelal gepleit voor het invoegen van zoveel als mogelijk variabelen, bijvoorbeeld van "het aantal dagen in het ziekenhuis" als indicator van morbiditeit. Er is besloten om medisch aanbod niet in de REF vergelijking op te nemen en daarmee ziekenfondsen verantwoordelijk te maken voor de regionale kostenvariatie die door de aanwezigheid van medisch aanbod wordt veroorzaakt, hoewel ziekenfondsen geen geschikte instrumenten hebben om de kosten van hun verzekerden te beïnvloeden (Van de Ven et al. 2007). Het in deze studie ontwikkelde theoretisch raamwerk kan worden gebruikt om het Belgische besluit om niet te compenseren voor regionale kostenvariatie nader te evalueren.

Sinds 2004 past het CMS in de VS een "broosheid" REF criterium toe ter financiering van PACE organisaties die in het kader van Medicare integrale diensten leveren aan groepen oudere mensen met functionele beperkingen die daardoor thuis kunnen blijven wonen in plaats van te moeten worden geïnstitutionaliseerd. Een "broosheid" criterium zal in 2008 niet programmabreed aan de CMS-HCC REF vergelijking worden toegevoegd ten behoeve van Medicare Advantage verzekeringen

vanwege verscheidene methodologische problemen die samenhangen met het gebruik van enquêtegegevens om de risicoafhankelijke premiesubsidies te berekenen (CMS 2007, Bijlage II, Sectie A). Echter, CMS kondigt aan dat zij door zullen gaan met onderzoek naar invoering van factoren in de CMS-HCC REF vergelijking zodat kosten die voortvloeien uit de "broosheid" van individuele verzekerden beter worden voorspeld. Uit de resultaten van deze studie blijkt dat gegevens over gebruik van fysiotherapie en medische hulpmiddelen kunnen worden gebruikt om de methodologische problemen waarvoor het CMS zich ziet gesteld te overwinnen. In de tussentijd zouden de REF gewichten van de huidige REF criteria alvast kunnen worden aangepast zodanig dat betere kruissubsidies voor de S-type kostenvariatie vanwege functionele beperkingen resulteren.

Beperkingen van deze studie

Het gegevensbestand dat in deze studie is gebruikt om de toepassing van het theoretisch raamwerk te illustreren, heeft betrekking op ziekenfondsverzekerden die zowel in 2001 als in 2002 bij Agis Zorgverzekeringen verzekerd waren. De hier gepresenteerde resultaten zijn daarom niet representatief voor alle geografische regio's in Nederland en ook niet voor voormalig particulier verzekerden waarvoor het REF model sinds de Nederlandse Zorgverzekeringswet van 2006 ook geldt. De kruissubsidies die zijn gebaseerd op de REF criteria van 2004 zijn niet berekend voor verzekerden jonger dan 16 jaar, omdat de Agis Gezondheidsenquête 2001 niet onder deze populatie is uitgezet.

De specificatie van de Nederlandse REF vergelijking uit 2004 verschilt in een aantal opzichten van de vormgeving waarvoor in deze studie is gekozen. Het REF criterium leeftijd is ingedeeld in tienjaarsklassen in plaats van vijfjaarsklassen en er is afgezien van interacties tussen leeftijd en verzekeringsgrond in deze studie. De kostendefinitie die in deze studie wordt gehanteerd bevat de vaste kosten ziekenhuisverpleging, terwijl deze kosten in de Nederlandse praktijk tot 2006 voor 95% werden nagecalculeerd. Sinds 2006 wordt ongeveer een derde van de ziekenhuiskosten als vaste kosten ziekenhuisverpleging gedefinieerd en voor 100% nagecalculeerd. Ten slotte worden de systemen van ex-post generieke verevening (tussen verzekeraars onderling) en nacalculatie (tussen verzekeraars en de sponsor) die in 2004 gelden ten aanzien van de variabele kosten van ziekenhuisverpleging en specialistische hulp niet in deze studie toegepast.

Het is mogelijk dat de S-type criteria die in deze studie in de normatieve vergelijking zijn opgenomen niet volledige de kostenvariatie beschrijven zoals die door de S-type risicofactoren worden bepaald. De in deze studie gekozen reeks van gezondheidsindicatoren kan in de toekomst worden uitgebreid, afhankelijk van de beschikbaarheid. De crux bij de toepassing van het in deze studie ontwikkelde

theoretische raamwerk is dat de gekozen reeks van S-type criteria minder beperkt is dan de beschikbare verzameling van REF criteria die in de praktijk wordt gehanteerd. In deze zin geeft de hier ontwikkelde methode een benedengrens aan voor de mate waarin de REF vergelijking de beoogde kruissubsidies genereren. Met andere woorden, gegeven de in deze studie gehanteerde operationalisatie van de normatieve kosten, zullen de prestatieindicatoren van de REF modellen zoals berekend in de hoofdstukken 6, 7 en 8 vormen naar verwachting een maximum aangeven van de mate waarin de kruissubsidies aansluiten op de beoogde doelen van de Nederlandse overheid.

De bijdrage van de procedure van weggelaten variabelen om de vertekening van de imperfecte REF gewichten te verwijderen blijkt nogal beperkt, uitgaande van de verzameling N-type criteria die in deze studie zijn toegepast. Deze resultaten kunnen veranderen als een andere, mogelijk uitgebreidere verzameling van N-type criteria zou worden toegepast.

Vervolgonderzoek

De in deze studie gepresenteerde empirische resultaten hebben betrekking op de populatie van Agis ziekenfondsverzekerden in 2002. Hoewel niet wordt verwacht dat de REF gewichten significant zullen veranderen voor de meeste REF criteria, zal dit onderzoek moeten worden herhaald voor de totale populatie van ziekenfondsverzekerden om in ieder geval in regionaal opzicht representatief te kunnen worden genoemd. In het bijzonder zouden de regionale REF gewichten hierdoor kunnen veranderen.

Verder is het zo dat het Nederlandse REF model sinds de Zorgverzekeringswet van 2006 betrekking heeft op meer dan 16 miljoen Nederlandse inwoners in plaats van de 10 miljoen ziekenfondsverzekerden. Het in deze studie ontwikkelde theoretisch raamwerk zou daarom op deze totale Nederlandse populatie moeten worden toegepast. De normatieve vergelijking kan in dat geval worden bepaald op basis van informatie uit nationale enquêtes die al beschikbaar zijn (bijvoorbeeld de module 'Gezondheid' in het POLS van het CBS).

In een competitieve zorgverzekeringsmarkt zonder risicoverevening zullen premiekortingen door zelfselectie bij vrijwillige eigen risico's voor een deel voortkomen uit kostenvariatie die door S-type risicofactoren wordt veroorzaakt (Van Kleef, Van de Ven en Van Vliet 2006). Dit zal tot op zekere hoogte nog steeds het geval zijn als de kruissubsidies op incomplete en/of imperfecte REF criteria zijn gebaseerd. De mate waarin hiervan sprake is kan worden bepaald door de in deze studie ontwikkelde aanpak. Verder kan het niveau van het vrijwillige eigen risico als proxy voor gezondheid in de REF vergelijking worden opgenomen. Hoewel toepassing van

deze proxy als nieuw REF criterium tevens tot niet beoogde kruissubsidies voor N-type kostenvariatie kan leiden, kunnen deze effecten binnen het in deze studie ontwikkelde theoretisch raamwerk expliciet tegen elkaar worden afgewogen. Een aanpassing van het desbetreffende REF gewicht kan eventueel worden uitgevoerd om de compensatie voor deze N-type effecten te voorkomen. Merk op dat een soortgelijke aanpak al in het empirische deel van deze studie is uitgevoerd met betrekking tot verzekeringsgrond en regio.

Het ultieme doel van gereguleerde concurrentie is dat verzekeraars hun rol als prudente zorginkopers oppakken. Om zover te komen zullen consumenten niet alleen prijsgevoelig zijn en van verzekeraar moeten willen veranderen, maar consumenten moeten ook in staat worden gesteld om kwaliteitsverschillen te kunnen waarnemen en ook hiervoor gevoelig te zijn en van verzekeraar willen veranderen. Onder druk van marktwerking zullen verzekeraars aldus worden gedwongen om zich bij het organiseren en inkopen van de zorg te richten op de voorkeuren van hun verzekerden. Een meer directe manier om te stimuleren dat verzekeraars inspelen op de voorkeuren van hun verzekerden is door de kruissubsidies af te laten hangen van expliciete indicatoren van de kwaliteit van gecontracteerde zorg (in aanvulling op de compensatie voor S-type kostenvariatie). Het IOM (2006) adviseert om een financieringspool te creëren door middel van een reductie van de basisuitkering in het kader van Medicare voor iedere klasse van zorgaanbieders (ziekenhuizen, professionele verpleegfuncties, Medicare Advantage verzekeringen, dialysefaciliteiten, thuiszorg organisaties en medisch specialisten). In eerste instantie zouden de pay-for-performance programma's zodanig moeten worden opgezet dat zorgaanbieders worden beloond als zij goede prestaties leveren en hun prestaties in de loop der tijd significant weten te verbeteren. Gegeven dat deze indicatoren vaak alleen voor een beperkt aantal consumenten beschikbaar zullen zijn, kan een natuurlijke aanpak eruit bestaan dat deze indicatoren in de normatieve vergelijking worden opgenomen en via aanpassing van de REF gewichten tot uitdrukking worden gebracht.

Een goed gedocumenteerd geval van tekortschietende levering van noodzakelijke zorg aan sociale subgroepen in een samenleving kan in Shmueli (2000) worden gevonden ten aanzien van de Arabische verzekerden die in Israël wonen. Tabel A8.3 in deze studie laat zien dat in Nederland de REF voorspelde kosten voor allochtonen van de eerste generatie enigszins boven de werkelijke kosten liggen, maar tegelijkertijd blijkt dat REF voorspelde kosten significant lager uitvallen dan de normatieve kosten. Dit betekent dat er sprake is van ernstig ondergebruik door allochtonen van de eerste generatie. Omdat REF verwachte kosten nauwelijks verschilt van de werkelijke kosten, is er voor verzekeraars vanuit financieel oogpunt

geen financiële prikkel om eventueel ondergebruik bij allochtonen van de eerste generatie weg te nemen, omdat dit tot voorspelbare verliezen zal leiden. Hoewel dit probleem volgens Schokkaert et al. (2006) wellicht moet worden aangepakt met directe subsidies of scholingsprogramma's, kan een alternatieve aanpak zijn om allochtonen van de eerste generatie als REF criterium in de REF vergelijking op te nemen en tegelijkertijd het desbetreffende REF gewicht voor N-type kostenvariatie te corrigeren. Deze aanpak maakt allochtonen van de eerste generatie tot goede risico's voor verzekeraars. Merk overigens op dat verzekeraars dan tegelijkertijd in staat moeten worden gesteld om de hoogte van hun premies ten aanzien van deze subgroep te differentiëren, om het gevaar te verminderen dat de hiermee ontstane voorspelbare winsten voor andere zaken dan het tegengaan van het ondergebruik van deze groep zal worden aangewend door de verzekeraars.

Samenvattend kan worden gesteld dat de benadering van risicoverevening in deze studie op verschillende manieren in de praktijk kan worden toegepast en relevant is voor alle landen met competitieve markten voor zorgverzekeringen. Er wordt in deze gevallen aanbevolen om de kruissubsidies van de REF modellen aan de hand van deze aanpak te toetsen en te verbeteren.

CURRICULUM VITAE

Pieter J.A. Stam was born and raised in Schoonhoven, The Netherlands. At secondary school "De Drie Waarden" in Schoonhoven he did exams in mathematics, economics, history, Dutch and the foreign languages English, French and German. He studied econometrics at Erasmus University Rotterdam, during which he also worked as a research assistant to prof. dr. B.M.S. Van Praag. After obtaining his masters degree, he worked on a broad range of consultancy projects at SEO Foundation for Economic Research at the University of Amsterdam for two years. In 1996, he started working in the health insurance sector at Agis Health Insurance (formerly known as ZAO Health Insurance).

Since 1998, he is an active member of all relevant working groups and committees on risk equalization modeling in the Dutch health insurance sector. He also belongs to the group of people who were the first to develop and implement the Dutch version of the CAHPS questionnaire in Dutch health care (renamed to "CQ-Index") and the so-called "pay-for-performance" strategy at Agis. He is member of the Scientific Board of the Miletus Foundation that safeguards the usefulness of the CQ-Index for the Dutch health insurers Agis, Menzis, VGZ and Delta Lloyd.

Since 2000 he has an affiliation with the institute for Health Policy and Management (iBMG) at Erasmus University Rotterdam and since 2003 he is member of the European Risk Adjustment Network (RAN) for health scientists that work in the field of competitive health insurance markets and risk equalization modeling. In 2004 he set up the risk equalization model at the individual member level to facilitate Agis to quantify their business cases for targeted groups of enrollees. This novum amongst Dutch health insurers is now part of regular business at Agis.

His national and international publications thus far are on discrete choice and conjoint analysis techniques, poverty lines and equivalence scales, cost-benefit analysis of sporting injuries, preferences of visitors for classical musical performances, preferences of high school graduates for university studies, patient experiences in health care (e.g. hospitals and diabetes care providers) and health insurance as measured by the CAHPS methodology, preferences and choices for quality in health care and health insurance, incentives for risk selection in Dutch basic benefits health insurance and incentives for risk-rating premiums in Dutch supplementary health insurance.

Ack.

Acknowledgments

I want to thank Ria for her love and support during the years. Every now and then, you offered me the time and space to pay attention to my research activities and nothing else, which came at the expense of the time that we were able to spend together. Furthermore, I want to thank my family and friends for their warmth and attention. In particular, I cannot thank my mother enough for her love and never ending support at each and every stage of her and my own life. It is an understatement to say that I miss you, since you passed away in July this year. I regret you can not be present at the defense of my PhD thesis, but most of all I miss you at sharing the more important little things in life.

The research activities for my PhD thesis were undertaken at the location of the Institute of Health Policy and Management under the authority of the Erasmus University Rotterdam. Wynand van de Ven and René van Vliet supervised the study, for which I am grateful. I thank you for your close reading of the draft versions, critical comments and enormous enthusiasm concerning the issues of regulated competition in the health care sector. Thanks also to the members of the Risk Adjustment Network (RAN) for adding an international perspective to these issues and for the stimulating discussions in general.

I want to thank my colleagues at Agis Health Insurance and the Institute of Health Policy and Management for providing a nice working environment. In particular, I want to thank Paul Schmidt for teaching me the basics of risk equalization when I started working in this field at ZAO Health Insurance, and Xander Koolman for the joyful "econometric" discussions about almost every subject one can imagine. Thanks both of you for being very critical and constructive when discussing the draft versions of my thesis.

I want to express my gratitude to the respondents to the Agis Health Survey 2001. Without their contribution this research project would not have been possible. Leida Lamers and Marleen Foets are acknowledged for their valuable input when selecting the survey questions. Agis Health Insurance, APE Public Economics and Prismant are acknowledged for providing data that were necessary to construct the risk equalization models in this study.

This PhD thesis is written in OpenOffice 1.1.4, an open-source alternative to the closed-source office suites, such as Microsoft Office, which was not available for my non-standard Agis laptop. I want to thank the OpenOffice.org developers and end-users for doing a marvelous job, as their software provided me with more than just a good alternative. I favor their mission.

Agis Health Insurance has financially supported the writing of this PhD thesis, treating it as one of their regular R&D projects. I am grateful for the opportunity they gave me to do scientific research two days a week during these years. In particular, I am indebted to Arnold van der Lee who managed to convince the Agis

board of directors that this research project could add value to their business. Furthermore, I thank you and Onno van der Galiën for always being inspiring when talking about daily life risk equalization and plenty of other topics. Note that the opinions expressed in this study are those of the author. No official endorsement of Agis Health Insurance is intended or should be inferred.