“LONG-TERM ISSUES

IN

HEALTH AND INSURANCE ECONOMICS”

Dissertation

for the Faculty of Economics, Business Administration

and Information Technology of the University of Zurich

to achieve the title of

Doctor of Economics

presented by

Patrick Eugster

from Meilen ZH

approved at the request of

Prof. Dr. P. Zweifel

Prof. Dr. D. Pfaff

The Faculty of Economics, Business Administration and Information Technology of the University of Zurich hereby authorises the printing of this Doctoral Thesis, without thereby giving any opinion on the views contained therein.

Zurich, 22.10.2008

the Dean: Prof. Dr. Dr. Josef Falkinger

Preface

This doctoral thesis has been written during my time at the Socioeconomic Institute (University of Zurich). From December 2003 until April 2008 I was able to work at the institute as a doctoral student and focus on writing my dissertation in economics. After completing my degree in economics at the University of Zurich in December 2003, I attended the program for beginning doctoral students in economics at the study center Gerzensee in 2004 and the beginning of 2005. I am grateful to Prof. Dr. Peter Zweifel for his ongoing support, all the other opportunities and for creating such a stimulating environment. It will always be a pleasure to remember the numerous and highly interesting projects and assignments which I carried out with him and how much I benefited from the scientific experience at his chair.
Moreover I want to thank Prof. Dr. Dieter Pfaff and Prof. Dr. Ulrike Stefani for the additional research questions I was able to work on with them. Even though these contributions are not part of this thesis the diversity of challenges were a great pleasure and I enjoyed them very much. In addition I want to thank Prof. Dr. Dieter Pfaff for evaluating and grading this thesis.
I would also like to mention Dr. Harry Telser and thank him for all the extensive discussions, the joint "cordon-bleus" and for organizing the illustrious circle of monthly economist-gatherings. In many difficult situations he motivated me and sent me back to work with new ideas. The one probably most often spoken to was Boris Krey. He too listened patiently to my ideas and argued with me about various issues many times. The time we spent together at the Socioeconomic Institute means a lot to me and will surely always be remembered. At this point I would also like to mention our two predecessors Dr. Karolin Becker and Dr. Yves Schneider. Karolin has introduced me excellently to the chair and my persistency in mathematical questions is definitely a result from working together with Yves. These four enabled me to dive into research, supported and challenged me in numerous discussions. PD

Dr. Konstantin Beck I want to thank especially for his support and for providing data in developing chapter five and for all the entertaining joint conference attendances. I would also like to mention all the other regular and irregular guests at the monthly economist-gatherings and thank them for the discussions about and beyond economics.
Of course I would also like to thank my former work colleagues: Michèle Sennhauser, Angelika Braendle, Maria Trottmann, Maurus Rischatsch, Johannes Regenass, Johannes Schoder, Phillip Widmer und Ilija Neustadt. The various bets on prices of equities, exchange rates, soccer games, European and World Championships, and never to forget Swiss beauty contests often had me come to work with great joy. Last but not least I want to thank the whole team of the library for their everlasting patience with renewals, transcriptions and various other small things. It was always a great pleasure seeing you on our excursions and other occasions.
This is also to my parents, Irene Eugster-Hnid and Hans-Rudolf Eugster, who I owe it to in great parts that I could enjoy such a long-lasting education. And of course also to my brother Michael Eugster, as well as Denise Laufer, Mischa Brennwald, Ralph Kistner, Fabian Peter, Rana Nag, Thomas Fothergill, Andreas Feurer and Claudio Piscitelli. They all endured my venture and often supported me in many ways. I want to thank Sarah Lord and Suzana Sulong Eugster for their English input reading my introduction and conclusion. Without the grand support from my environment I probably would have never finished this thesis. THANK YOU SO MUCH!

P.E., Zurich, April 2008

Contents

Vorwort **i**

1 Introduction **1**

2 The Sisyphus syndrome in health revisited **7**

2.1 Introduction 7

2.2 Survey of the literature 9

2.2.1 The production of health 9

2.2.2 The health care expenditure (HCE) function 11

2.2.3 A combined view: the Sisyphus syndrome 13

2.3 Specification, data base, and variables 15

2.4 Estimation results 19

2.4.1 Production of health 19

2.4.2 Explaining the Health Share in the GDP 22

2.4.3 Assessing the strength of the Sisyphus syndrome 25

2.5 Conclusion 27

3 Life-cycle effects of social security in an open economy:
A theoretical and empirical survey **33**

3.1 Introduction 33

3.2 Educational effort and social insurance 35

3.2.1 Financing of social security and educational effort 35

3.2.2 Social security benefits and educational effort 36

3.3 Marriage and social insurance 38

3.3.1 Financing of social security and marriage 38
3.3.2 Social security benefits and marriage 38
3.4 Number of children and social insurance 40
3.4.1 Financing of social security and number of children 40
3.4.2 Social security benefits and number of children 40
3.5 Divorce and social insurance 42
3.5.1 Financing of social security and divorce 42
3.5.2 Social security benefits and divorce 42
3.6 Retirement and social insurance 43
3.6.1 Financing of social security and retirement 43
3.6.2 Social security benefits and retirement 44
3.7 Life expectancy and social insurance 46
3.7.1 Financing of social security and life expectancy 46
3.7.2 Social security benefits and remaining life expectancy 46
3.8 Conclusion 48

4 Correlated risks: A conflict of interest between insurers and consumers and its resolution 57

4.1 Introduction 57
4.2 Multiple risks: theory and stylized facts 58
4.2.1 Multiple risks in insurance economics 58
4.2.2 How are assets of individuals correlated? 59
4.2.3 An application of portfolio theory 61
4.3 Data 63
4.4 Econometric specification 64
4.5 Analysis of short-term correlations 66
4.5.1 United States 66
4.5.2 Switzerland 69
4.6 Long-term analysis 72
4.7 Possibilities of international diversification 76
4.8 Conclusion 76

5 Capping risk adjustment? 87

5.1 Introduction . . . 87
5.2 Data basis, descriptive statistics, and representativeness . . . 88
5.3 Hospitalization as an additional criterion . . . 92
5.4 Limiting the volume of risk adjustment . . . 92
5.5 Optimal cap on the volume of risk adjustment . . . 96
5.6 Consequences of capping risk adjustment . . . 99
5.6.1 Empirical illustration . . . 101
5.7 Conclusion . . . 104

6 Can guaranteed renewability survive in the presence of death? 117

6.1 Introduction . . . 117
6.2 Related Literature . . . 118
6.3 The Model . . . 121
6.3.1 Positive probability of returning to favorable risk status . . . 123
6.3.2 Introducing Death I: Time of Death . . . 125
6.3.3 Introducing Death II: Extra Cost of Dying . . . 128
6.4 Credit Rationing Revisited . . . 128
6.5 A More General Approach . . . 131
6.5.1 Positive Rate of Interest and Increasing Health Care Cost . . . 131
6.5.2 Determining the Critical Value of Time Preference . . . 132
6.6 Conclusion . . . 137

7 Conclusion 147

Chapter 1

Introduction

This collection of essays explores questions related to long-term issues in insurance and health economics. However, the first five chapters cover heterogeneous topics and look at miscellaneous issues from differing points of view. Chapters 2, 4 and 5 develop and test empirical models and their predictions, while chapter 3 surveys the theoretical and empirical literature, and chapter 6 presents a theoretical model. Chapter 7 posits some policy implications and discloses possible future extensions to chapters 4, 5, and 6.

The so-called Sisyphus syndrome is the concern of chapter 2: Increased past health care expenditure is shown to increase the share of aged individuals which in turn results in additional current expense. Chapter 3 examines life-cycle effects of social security in an open economy. Surveying the theoretical and empirical literature, it seeks to endogenize demographic change. Long-term common trends in private insurance benefits are analyzed in chapter 4: Possible hedging strategies in and between both the United States and Switzerland are identified in the long term. In addition, short-term correlations between unexpected deviations of social and private insurance benefits are estimated.

The following two chapters address issues regarding risk selection in long-term health insurance. Taking up community-rated premiums, chapter 5 studies whether the volume of risk adjustment changes when its formula is complemented by an additional criterion and simulates a possible cap on the volume. Chapter 6 explores the concept of guaranteed renewability, which is a more market-based solution. Does it survive when death and a positive transition probability from high- to low-risk status are modelled?

"The Sisyphus syndrome in health revisited", chapter 2, notes that when the task is about to be completed, ironically work must start all over again. To see the analogy, consider an initial decision to allocate more resources to health. The likely consequence is an increased number, and prolonged lifespan, of survivors, who will exert additional demand for health care. With more resources allocated to health, the cycle recurs again. The objective of this chapter is to improve on earlier research that failed to find evidence of a Sisyphus syndrome in industrialized countries. This time, however, albeit signs of such a cycle exist, the effect is non explosive.

Both chapters 2 and 3 deal with a feedback process in a long-term setting. While chapter 2 looked at how additional health care expenditure increases remaining life expectancy and drives the process, chapter 3 examines the effect of old-age provision, health insurance, and long-term care insurance on decisions throughout one's lifetime. Demographic change is usually perceived as an exogenous shock impinging on social security, with the economy treated as a closed system. This chapter argues that demographics is the aggregate of individual decisions influenced by social security. This claim is supported by both theoretical argument and empirical evidence, with regard to decisions over the life cycle spanning educational effort, marriage, number of children, divorce, retirement, and effort to extend one's life. These feedback relationships are shown to hamper employment and growth, and thus to undermine the financial viability of today's social security schemes – with increasing openness of the economy ("globalization") exacerbating problems.

Chapters 4 to 6 survey long-term issues in the insurance industry, including varying insurance systems (community-rated and risk-rated systems). Chapter 4 looks at short- and long-term correlations of trend deviations in different lines of insurance and points to a conflict of interest between consumers and insurers. Consumers face positive correlation in their assets (health, wealth, and wisdom; in analogy to the theory of chapter 2, that is based on the same three assets, see Figure 2.1), causing them to demand a great deal of insurance coverage. Insurers, from the other standpoint, eschew positively correlated risks. It can be shown that insurance contributes to a reduction of consumers' asset volatility only if unexpected deviations of payments from expected value correlate negatively across lines of insurance. Analyzing deviations from trend in aggregate insurance payments, one finds that private U.S. (but not Swiss) insurance has a hedging effect for consumers, whereas both social insurance schemes expose consumers to excess asset volatility. In the insurance systems of both countries, the private component fails to offset deviations in the social component, and vice versa. As to the supply of insurance, cointegration analysis indicates the absence of common trends. Insurance companies could therefore offer combined policies to the benefit of consumers, hedging their underwriting risk both domestically and internationally.

One solution for dealing with risk selection in long-term health insurance is to impose community rating by law; the other is to make contracts incentive-compatible. Chapter 5 is devoted to the first and chapter 6 to the latter. When premiums are community-rated, risk adjustment serves to mitigate competitive insurers' incentive to select favorable risks. Based on age and sex as the sole criteria, risk adjustment in Switzerland has increased by 80 percent in volume (since its introduction in 1996), attaining as much as 31 percent of health insurers' total income. This development has triggered calls for capping risk adjustment. At the same time, the debate revolves around a refinement of the adjustment formula by factoring in hospitalization or living in a nursing home during the previous year. Inclusion of this criterion is found not only to increase the volume but also the distribution of risk adjustment between consumers. It also would increase the opportunity cost of capping risk adjustment in terms of increased incentives for risk selection.

Chapter 6 is titled "Can guaranteed renewability survive in the presence of death?" and is based on the seminal contribution by Pauly et al. (1995) regarding time consistency and guaranteed renewability (GR) in insurance contracts. GR denotes that both parties are always willing to continue the sequence of contracts even after the occurrence of a loss or a change in risk status. In return, the contract is frontloaded in order to generate the necessary reserves. However, frontloading can become excessive. This chapter adds to the debate by introducing the full set of relevant transition probabilities. Whereas Pauly et al. (1995) consider only the likelihood of a favorable risk turning into an unfavorable one during the planning period, both the possibility of returning to favorable risk status and death are accounted for here. Returning to favorable risk status can reduce the amount of frontloading in a model covering more than three periods – with the amount of reduction increasing with the probability of returning to favorable risk status. Moreover, it is shown that once death and the cost of dying are included in the model, then the critical value of time preference that is still compatible with GR decreases, causing demand for GR contracts to increase. In conclusion, once the underlying Markov model is completed, GR cannot be said to survive but to mitigate the problem of risk selection in health insurance to an even greater extent than previously claimed. Note that Peter Zweifel co-authored chapters 2-5, Lukas Steinmann co-authored chapter 2, and Michèle Sennhauser co-authored chapter 5. The undersigned author was at least equally responsible for the intellectual input of chapters 2, 4 and 5, while the principal contribution to chapter 3 is by Peter Zweifel. Chapter 2 appeared in the *International Journal of Health Care Finance and Economics*, and chapter 3 in the *Zeitschrift für die gesamte Versicherungswissenschaft*. Chapter 4 has been submitted to the *Journal of Risk Management in Financial Institutions*, chapter 5 to the *Journal of Health Economics*, and chapter 6 will be submitted to the *Journal of Risk and Uncertainty*.

Patrick Eugster
Zurich, March 2008

Bibliography

Pauly, M. V., H. Kunreuther, and R. Hirth (1995). Guaranteed renewability in insurance. *Journal of Risk and Uncertainty 10*(2), 143–156.

The Sisyphus Syndrome in Health Revisited

Peter Zweifel, Lukas Steinmann and Patrick Eugster*

Published in the "International Journal of Health Care Finance and Economics", 5(2), 2005, 127-145.

Chapter 2

The Sisyphus syndrome in health revisited

2.1 Introduction

When the British National Health Service was created, its founders expected the problem of ill health to fade away. After all, once people have access to good medical care, they should live healthy and happy to the end of their lives. History has proven this expectation to be entirely false. Indeed, decision makers in health policy are haunted by a quite different suspicion, which Zweifel and Ferrari (1992) dubbed the Sisyphus syndrome. Health care services could be something like the work of Sisyphus, a Greek hero who was ordered by the Gods to push a big rock uphill. But just before the very top, the rock would slip from his hands so he had to start all over again.

Consider the following sequence of events in health care. Initially, politicians decide to allocate more resources to health. If effective this intervention causes people that would have otherwise died to survive. With more survivors around, there will be additional demand for health care services. To the extent that this is financed out of private resources, there is not too much of a problem. Individuals will adjust their health insurance policies accordingly and allocate a greater share of their income to health care. However, most of these services are covered by public health insurance. Rather than accepting to pay themselves, especially older voters have an incentive to get politicians to reallocate the public budget in favor of health. The increase in health care expenditure (*HCE*) again creates survivors. Thus, the Sisyphus syndrome can go into its next turn.

For the Sisyphus syndrome to exist, two (possibly lagged) relationships must be operative.

1. Health care expenditure (*HCE*) must have the effect of increasing remaining life expectancy (*RLE*). For policy, this effect is of particular importance if it occurs at retirement age because in systems financed by a payroll tax, retired individuals do not contribute to the financing of health care anymore. In tax-based systems, their contribution decreases to the extent that their income drops after retirement.

2. Increasing *RLE* must translate (through a changed planning horizon and through an increased number of survivors) into an impact on health care spending.

However, the second relationship is confounded in that lower rather than higher *RLE* is sometimes associated with *HCE*. Evidence is accumulating that much of lifetime *HCE* is spent during the last year before death [Lubitz and Riley (1993); Zweifel et al. (1999)]. In this case, *HCE* does not generate survivors that will spend more private *HCE* and exert pressure for more public *HCE* in the future. At the macroeconomic level, the share of the population that is in their last year of life is reflected by the mortality rate. A high mortality rate (which itself may depend on previous *HCE*) is associated with high *HCE* in the current period while mitigating the dynamics of the Sisyphus syndrome.

As to relationship (1.), Zweifel and Ferrari (1992) found that prior *HCE* in OECD countries did increase *RLE* at ages 40 and 65. But with regard to relationship (2.), they were not able to establish a statistical link between *RLE* and either public or private *HCE*. The dampening effect of mortality was not accounted for. In all, a Sisyphus syndrome could not be said to exist.

The objective of this chapter is to revisit these findings. In the light of improved specification, data, and econometric methodology, what can be said about the effect of *HCE* on *RLE*, especially at higher ages? Is it still true that no evidence can be found of a feedback, from increased *RLE* to *HCE*? And is it necessary to control for the mortality rate? Accordingly, the plan of this chapter is as follows. In section 2.2, a review of the relevant literature is provided, followed by a more precise restatement of the Sisyphus syndrome. Section 2.3 is devoted to econometric specification, a description of the data base, and variable definitions. The new empirical evidence is presented in section 2.4, while section 2.5 contains a conclusion.

2.2 Survey of the literature

2.2.1 The production of health

The seminal article continues to be Auster et al. (1969), who related age- and sex-adjusted mortality rates of U.S. states to medical care inputs, schooling, income, and environmental variables. Schooling was negatively related to mortality, medical care inputs not consistently so, while higher income tended to cause an increase in mortality. Hadley (1982) refined this approach, using county groups as unit of analysis. Again, education was negatively related to mortality, income negatively to infant mortality but in mixed ways to adult mortality. A new finding is that most age-sex specific mortality rates depend negatively on medical care spending, with an elasticity of -0.15. Focusing on the elderly, Hadley (1988) found elasticities in the -0.25 to -0.44 range.

An early international study of the production of health is the one by Stewart (1971). Using the nations of the Western hemisphere in the mid-1960s as the unit of analysis, he related life expectancy at birth to medical inputs, literacy rates, and the availability of potable water. All three inputs had small marginal effects in the case of the United States, whereas literacy and potable water showed significant effects in the less developed countries of the group.

Cochrane et al. (1978) based their analysis on a cross-section of 18 developed countries as of 1970. In their correlation analysis, a greater number of physicians goes along with higher rather than lower mortality (with no attempt to control for reverse causation). Mortality rates correlate negatively with income per capita, suggesting that the positive effect found by Auster et al. (1969) might be limited to the United States. They also tested the effects of cigarette and alcohol consumption, with mixed effects. The ambiguous role of income may be due to failure to account for unfavorable life styles that accompany the achievement of higher (labor) income. This is the conclusion that can be drawn from Gerdtham and Ruhm (2006), who find that out of 9 specific causes of mortality in OECD countries, 6 tend to be negatively related to income, while mortality from heart disease displays a strong and significant positive relationship. At the same time, there is evidence in 5 out of 9 cases that a higher degree of demand pressure in the economy (proxied by a lower unemployment rate) may drive up mortality rates.

Working with a 1980 cross section of OECD countries, Zweifel and Ferrari (1992) introduced lagged *HCE* (the lag being 10 years due to data availability) as an explanatory variable on the grounds that earlier medical interventions have assured the survival of individuals that make up the current population. Also they replaced mortality by remaining life expectancy (at ages 40 and 65) to better represent the planning situation of forward-looking individuals. Lagged *HCE* proved significant, with an elasticity (at the mean) of 0.15.

More recently Frech III and Miller (1999) review research performed by noneconomists (focused on low-income countries), which tends to support the notion that literacy and diet but not medical care resources are determinants of health. They analyze an OECD cross section, in the main making remaining life expectancy (*RLE*) at ages 40 and 60 as of 1993 the dependent variable. The explanatory variables include *GDP*, the consumption of tobacco and alcohol, and the consumption of animal fats. The authors introduce lags of up to ten years (depending on data availability) to capture lifetime effects because variations in most of these variables are unlikely to affect health right away. Rather than considering total *HCE* only, they single out pharmaceutical consumption to find that this component matters far more than does total *HCE*, which fails to have a significant effect on *RLE*. In their analysis of a cross section for the year 2000, Miller and Frech III (2004) complement their set of explanatory variables with obesity (which proves significant with the expected sign) but continue to find that pharmaceutical consumption importantly contributes to both unadjusted and disability adjusted *RLE* at age 40 and 60. Their specification is double logarithmic throughout, resulting in the estimation of constant elasticities which necessarily imply decreasing marginal productivity of *HCE* as soon as *HCE* differs more strongly between countries than does life expectancy (both relative to a benchmark, e.g. the United States).

Although this review may not be complete, the paucity of studies on the production of health is striking. By way of contrast, there has been much more research into the determinants of *HCE* (to be reviewed in section 2.2.2 below). Indeed, one might argue that the separate estimation of a production function is not necessary, based on a duality argument. If individuals acted as perfect cost minimizers in their production of health, the cost function [or an expenditure share function prior to normalization of some price, cf. e.g. Christensen et al. (1973)]

contains all the information relating to the production technology. However, the estimation of a separate production function of health can be justified by the following arguments:

1. Beyond differentiating between medical and broadly economic explanatory variables, it is still not clear which inputs belong in the health production function and whether their marginal productivities are stable across countries and across time.[1]

2. The health production technology is stochastic rather than deterministic. The deployment of both medical and nonmedical inputs does not always prevent death, i.e. a health outcome that differs strongly from the one intended. This also implies that cost minimization is possible only on expectation, with large deviations that serve to qualify the duality argument.

3. In the case of health, prices contained in the cost or expenditure share function are average prices prior to insurance and public subsidies, while prices governing individual cost minimization are marginal prices often close to zero and different from average prices.

4. Cost minimization is not a very credible hypothesis in health since decisions frequently are made under the influence of physicians. A distortion results to the extent that physicians act as imperfect agents of their patients. Moreover, *HCE* at the aggregate level may well reflect political institutions.

Conclusion 1. *The estimation of a production function for health apart from a cost or expenditure (share) function can be justified. The emerging consensus seems to be that at least some components of health care inputs contribute to health with a lag and decreasing marginal effect.*

2.2.2 The health care expenditure (HCE) function

Kleiman (1974) was one of the first to relate *HCE* to *GDP* using a cross-section of OECD country data. He also included the population share of those aged 64+ but was unable to find a statistically significant effect. Accordingly, Newhouse (1977) focused on the relationship between per capita *GDP* and *HCE*, using a cross-section of OECD countries. The computed

[1] We owe this important point to one of the referees.

income elasticity indicated that health care is a luxury good. This finding was confirmed by Maxwell (1981) in a broad international survey.

In another regression analysis based on OECD data, Leu (1986) reintroduced the share of persons aged less than 15 years and more than 65 years. However, his main innovation was the argument that a greater share of public in total *HCE* facilitates budget maximization by bureaucrats, which can be restrained by direct democracy or through central control, as in countries with a National Health Service. These predictions were confirmed but failed to be replicated by later studies.

In their first attempt at identifying the feedback from remaining life expectancy *RLE* to *HCE* as part of the Sisyphus syndrome, Zweifel and Ferrari (1992) found *RLE* at age 65+ to be even negatively related to *HCE*, although without statistical significance. Public *HCE* turned out to be as much of a luxury good as private *HCE*, with income elasticities far above unity. In a recent analysis of 70 year old Americans, Lubitz et al. (2003) again fail to find a relationship between differences in *RLE* (indicated by the presence of functional limitations) and projected future *HCE*. However, as will be argued in 2.2.3 below, it is sufficient for such a relationship to hold in a subset of the population in order to fuel the Sisyphus syndrome.

Collating information from three years to form an OECD panel data set for the first time, Gerdtham et al. (1992) were able to introduce several new explanatory variables, such as physician density, the share of inpatient to total *HCE*, as well as dummies for fee-for-service payment and budgeting caps. In the present context, it is interesting to note that a 10 percent increase in the share of inhabitants aged 64+ (relative to the 15 to 64 age group) was estimated to increase *HCE* by almost 2 percent, in contradistinction to all studies cited so far.

Extending the period covered to 1970-1991, Gerdtham et al. (1998) introduced a whole host of additional institutional variables. Among them, public reimbursement, capitation of physicians, and patient-first payment with later reimbursement consistently are associated with lower *HCE*. Tobacco consumption proves to be a significant predictor of higher *HCE*. With regard to the share of the elderly population (75+ this time), there is no indication of a significant effect (negative if anything).

A recent concern has been the possibility of spurious regression results due to nonstationarity of time series, in particular *HCE* and *GDP*. Without directly addressing the nonstationarity issue, Barros (1998) related the growth of *HCE* from 1960 to 1990 in OECD countries to the growth of *GDP*. The dummy variables representing gatekeeping systems, public reimbursement, and integrated systems, respectively, all prove insignificant once initial conditions are controlled for. Implied income elasticities range between 0.62 and 0.92. In their survey, Gerdtham and Jönsson (2000) report very mixed results of studies applying tests designed to detect nonstationarity. Since the emphasis of this chapter is not on the estimation of the income elasticity of *HCE*, this route is not pursued. Rather detrending is sought by transforming the dependent variable into the health share in *GDP*, resulting in the expenditure share function suggested by duality (normalization by a price should be avoided in this context, see section 2.3). In addition, public debate revolves much more about the share of *GDP* claimed by *HCE* rather than about *HCE* per se.

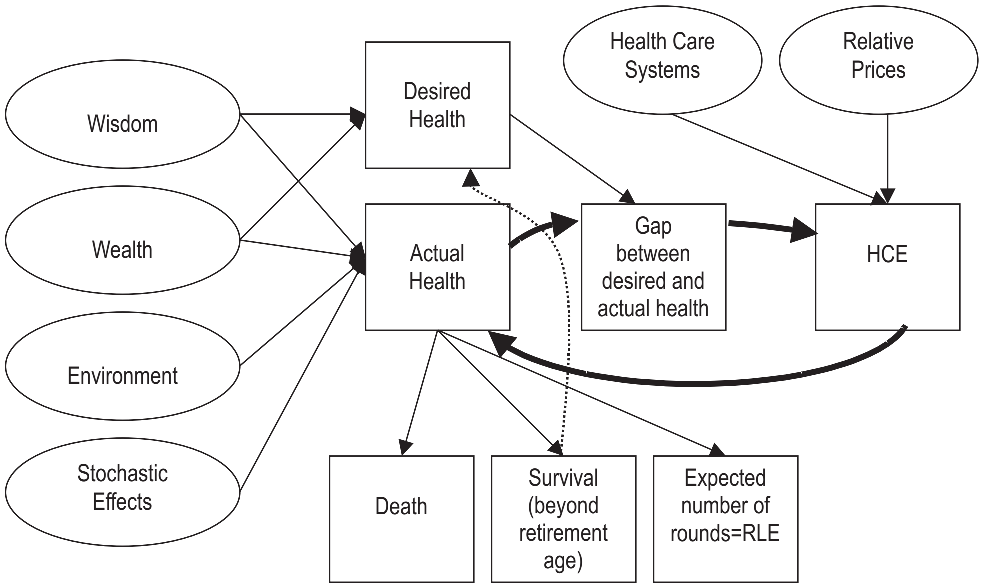
Conclusion 2. *With one exception [Gerdtham et al. (1992)], there is no evidence to the effect that a high share of individuals at or in retirement age leads to higher HCE. Other consistently significant regressors are GDP and likely tobacco consumption.*

2.2.3 A combined view: the Sisyphus syndrome

The theoretical starting point of this study is the fact that an individual not only has to manage his or her health stock but also a stock of wealth and a stock of human capital ("wisdom", see Figure 2.1). The dynamic optimization problem amounts to a series of portfolio choices determining the shares held in the guise of wisdom, wealth, and health. Formal modelling has concentrated on the health and wealth components of this portfolio [see Grossman (2000) for a comprehensive review]. Certainly with regard to the health component, this lifetime asset management is conditioned by environmental influences (e.g. degree of industrialization, degree of urbanization). The left-hand portion of Figure 2.1 thus shows the production part of the Sisyphus model.

The right-hand side of Figure 2.1 contains the derived demand part of the model. *HCE* is triggered by a gap between desired and actual health stock and influenced by the institutional characteristics of the health care system and relative prices. This implies that *HCE* is an endogenous determinant of health, a fact that has not been duly recognized in those existing studies on the production of health that introduce contemporaneous *HCE* as an exogenous explanatory variable. Next, the input of *HCE* increases (possibly with a lag) actual health

Figure 2.1: Sisyphus Syndrome



stock. If actual health stock remains below a critical level, death occurs, and the Sisyphus cycle comes to an end. If however the stock is sufficient to ensure survival, the individual concerned can count on performing the cycle for a number of periods. The expected number of future rounds defines his or her remaining life expectancy (*RLE*).

Thus, the feedback from *HCE* to actual health stock determines the number of survivors who are able to exert demand for health care and hence cause *HCE*. It is this feedback that is responsible for what can be called the Sisyphus syndrome: Given that *HCE* is increased exogenously (e.g. by an expansion of public health insurance coverage), more individuals survive the cycle. Moreover, these individuals have a longer planning horizon thanks to an increased *RLE*. This also means that the optimal health stock should have a higher value, which by itself makes the gap between desired and actual health stock larger, thus calling for additional *HCE* again (see dashed arrow in Figure 2.1). On the other hand, the direct effect of *HCE* is to close the gap between desired and actual health stock. Therefore, the net effect of *HCE* on the size of this gap remains an open issue, to be examined empirically.

This model motivates the creation of an explanatory variable that has not been introduced in previous work. It will be called *SISYPH* because it reflects the potential strength of the Sisyphus syndrome. *SISYPH* is the product of two factors: The first is *RLE* which shows the number of cycles an individual of a given age expects to perform. However, if the relative number of such individuals in the population is small, *RLE* by itself will have a small impact on the whole system. Therefore, *RLE* needs to be weighted by the share of the population at that age as the second factor. Since individuals at retirement age cannot be expected to contribute much to the financing of future *HCE* (creating a budgetary problem for policy makers), the share of those aged 65 and older will be used in the following.

This description of the Sisyphus syndrome is still not complete, however. Some individuals, having exerted demand for *HCE* on the expectation to survive for at least one more period, in fact die. Since the gap between desired and actual health stock likely is large in this event, these individuals will have driven up *HCE* in the current period while not contributing to the Sisyphus syndrome in the future. Therefore, this confounding effect (discussed in section 2.1) must be taken into account in order to avoid the risk of overestimating the dynamics of the syndrome. The share of deathbound individuals in a population is given by the mortality rate, *MORT*. This constitutes another explanatory variable that has been disregarded in published work on *HCE*.

2.3 Specification, data base, and variables

The theoretical considerations laid out in the preceding section provide guidance as to the specification of the relationships constituting the Sisyphus syndrome. Ideally, the hypothesis should be tested using individual panel data; however, data availability dictates the use of aggregate indicators.

In the present situation, two issues need to be clarified. One is choosing the right lags. It turned out that if both *GDP* and *HCE* have the same lag, multicollinearity becomes severe. Since the main focus of this chapter is on the effect of *HCE*, *GDP* is interpreted as a budget constraint rather than a lifestyle variable in the following, suggesting a zero lag, while *HCE*, alcohol consumption, and calorie intake are subjected to different lags. Preliminary estimates showed tobacco consumption to be insignificant, likely because the data base contains too many gaps with respect to this variable. The lags tested for are 3, 5, and 10 years, causing the sample to vary both in terms of countries and years covered due to missing values. A Hausman (1978) test is applied to select the preferred lag. It must be admitted of course that any lags exceeding 10 years cannot be discovered in this way.

The other issue is the choice of a random effect vs. fixed effect specification. The advantage of the random effects specification is that it preserves degrees of freedom; on the other hand, the country-specific component of the error term has to be independent of the explanatory variables for consistency. Fixed effects provide more information about the peculiarity of a country, but with 30 country-specific dummy variables, multicollinearity quickly becomes a problem. One solution is to drop insignificant dummy variables, which however entails a possible type II error each time. The regression strategy adopted here therefore is to estimate both variants and to decide again on the basis of a Hausman (1978) test.

In view of the uncertainty surrounding the production of health, a specification permitting variable marginal effects is retained for the first part of the Sisyphus model. This consists of Equations for remaining life expectancy (*RLEF*, *RLEM*) at age 60, the population share of individuals aged 65 and more (*POP65*, not available according to gender), and the mortality rate of the population as a whole (*MORT*). Subscripts i (countries) and t (year) are dropped for simplicity except for the error terms and $k = 3, 5, 10$. In the case of random effects, one has,

$$\begin{bmatrix} RLEF \\ RLEM \\ POP65 \\ MORT \end{bmatrix} = \begin{array}{l} \alpha + \beta_1 GDP + \beta_2 GDP2 + \beta_3 HCE_{-k} + \beta_4 HCE2_{-k} \\ +\beta_5 ALC_{-k} + \beta_6 ALC2_{-k} + \beta_7 CAL_{-k} + \beta_8 CAL2_{-k} \\ +\upsilon_i + \varepsilon_{it} \end{array} \tag{2.1}$$

Since these Equations have the same explanatory variables, the $\beta_1 ... \beta_8$ denote vectors containing four elements. The same holds true of υ_i and ε_{it} which describe a two-factor random effects specification, with no correlation across the four Equations assumed.

In the fixed effects alternative, the United States serves as a benchmark. The maximum of 29 dummy variables, D_i is not attained because of multicollinearity. Interestingly, the same 10 countries had to be deleted from the four Equations, yet despite this similarity between countries excluded, it proved impossible to benefit from the similarity of those retained. Attempts to group countries with a common dummy variable were not successful. The fixed effects Equation thus is defined as follows,

$$\begin{bmatrix} RLEF \\ RLEM \\ POP65 \\ MORT \end{bmatrix} = \begin{array}{l} \alpha + \beta_1 GDP + \beta_2 GDP2 + \beta_3 HCE_{-k} + \beta_4 HCE2_{-k} \\ +\beta_5 ALC_{-k} + \beta_6 ALC2_{-k} + \beta_7 CAL_{-k} + \beta_8 CAL2_{-k} + \gamma' D_i \\ +\mu_{it} \end{array} \tag{2.2}$$

The variables are defined as follows (see Table 2.1 for a statistical description) with $k = 3, 5, 10$,

- *RLEF*: Remaining life expectancy of females at age 60
- *RLEM*: Remaining life expectancy of males at age 60
- *POP65*: Population share of individuals aged 65 and more
- *MORT*: Mortality rate of the population, both genders, all ages, deceased per 100,000 inhabitants
- *HCE/GDP*: Share of health care expenditure in *GDP*, both nominal, in percent
- ALC_{-k}: Alcohol consumption per capita, in litres, lagged k years; $ALC2_{-k}$: ALC_{-k} squared
- CAL_{-k}: consumption of kilocalories per capita, lagged k years; $CAL2_{-k}$: CAL_{-k} squared
- *GDP*: *GDP* per capita, nominal but converted in US$ 1,000, *GDP2*: *GDP* squared
- HCE_{-k}: Total health care expenditure per capita, nominal but converted in US$ 1,000, lagged k years; $HCE2_{-k}$: HCE_{-k} squared
- *RPH*: Price index of health care services relative to price index of *GDP*
- *SISYPH*: Predicted *RLE*, averaged over gender, multiplied by *POP65*: [for details, see Equation (2.3)].

The source of all these data is OECD (2001), a source known for several difficulties. One of them is national differences with regard to the delimitation of the health care sector, resulting in different baskets of services, another, the lack of comparability and precision of health care

Table 2.1: Descriptive statistics of variables, 1970-2000

	Overall mean	1970 min[a]	1970 max[a]	2000 min[a]	2000 max[a]
$RLEF$	21.973	16.0	20.9	18.2	22.9
$RLEM$	17.678	14.1	17.8	16.0	21.4
$POP65$	13.369	3.1	14.1	5.3[b]	17.8[b]
$MORT$	845.888	852.3	1328.5	670.9[b]	1173.5[b]
HCE/GDP	7.571	2.584	6.951	5.412[b]	12.865[b]
RPH	89.820	54.8	121.6	98.6	108.1
GDP	15.762	0.272	5.027	2.974	42.315
HCE_{-10}	0.522	0.006	0.144	0.082	2.820
ALC_{-10}	11.018	0.9	23.7	1.4	16.6
CAL_{-10}	3.048	2.096	3.467	1.870	3.705
$\widehat{SYSIPH}$[c]	272.165				
N	233				

[a]The minima and maxima for 1970 and 2000 were taken over the whole sample available even if the observation was not used in the estimation of Table 2.5.

[b]1999

[c]Estimated variable, see Equation (2.3) in the text. Values for 1970 and 2000 not available due to missing values of either $\widehat{POP}65$, $\widehat{RLEF}$ or $\widehat{RLEM}$.

deflators. In view of these difficulties, *HCE* was not deflated using national price indexes but by the exchange rate when converting the figures into US$ [avoiding PPP indicators that may contain additional measurement error; see Gerdtham and Jönsson (1991)].

The variable *SISYPH* is constructed using predicted values from Equation (2.1) or Equation (2.2) as follows,

$$\widehat{SISYPH} = (\widehat{RLEF} + \widehat{RLEM})/2 * \widehat{POP65} \quad (2.3)$$

Finally, the derived demand part of the model is specified as an expenditure share, in keeping with the duality considerations of section 2.2.1. Thus, the core Equation for the health share in the GDP reads as follows with the correct error term added and the dummy variables if needed according to the model in use,

$$\begin{aligned} HCE/GDP &= \alpha' + \beta_1' GDP + \beta_2' RPH + \beta_3' \widehat{MORT} \\ &\quad + \beta_5' \widehat{SISYPH} + \ error \end{aligned} \quad (2.4)$$

Again, no correlation of error terms with those of Equations (2.1) and (2.2) is admitted. In contradistinction to Equations (2.1) and (2.2), these Equations do not contain squared terms anymore to avoid extreme multicollinearity. The variable $\widehat{SISYPH}$ already depends on all values of squared regressors [see Equation (2.1) and Equation (2.2)], while for RPH there is no theoretical reason to enter a squared value when the dependent variable is interpreted as a cost share Equation.

2.4 Estimation results

2.4.1 Production of health

As stated at the beginning of section 2.3, Hausman specification tests were performed to establish both a preferred error structure and choice of lag. In addition, Arellano-Bond tests were used to determine the order of average autocovariance in residuals. Fixed effects (FE) versus Random effects (RE) matched with lags of (3, 5, 10) years makes for 6 elements that are pitted against each other (a combination without replacement), or $\frac{6!}{(6-2)!2!} = 15$ possibilities. This totals 15 combinations where each such combination is to take FE as the maintained hypothesis and to test the RE against it.

It turns out that in all of the 15 comparisons, the FE alternative never dominates without an exception of at least one lag in at least one Equation making up (2.1) or (2.2). By way of contrast, the RE specification dominates without exception in three of these 15 paired tests, and in an additional three with no more than one exception. On the other hand when the Hausman test is applied to the three lags, (3, 5, 10) dominate in (6, 2, 7) comparisons. Both lags 3 and 10 have five exceptions, but in lag 3, they are multiple in one case. Therefore, the evidence suggests a random effects specification with a lag of $k = 10$ years.

The influence of *HCE* on Remaining Life Expectancy (*RLE*)

In the original article by Zweifel and Ferrari (1992), remaining life expectancy of males and females at ages 40 and 65 was distinguished, the main motivation being the quadrupling of observations. However, the differences between ages 40 and 60 in the present work proved minimal when the critical values were computed where additional HCE does not affect RLE

at the margin. As the policy focus is on individuals approaching retirement, only gender-specific results for age 60 are shown in Table 2.2.

Table 2.2: Remaining life expectancy at age 60, 1970-2000

	Females			Males		
RLE^a	Coefficient	z	$P > z$	Coefficient	z	$P > z$
GDP	0.122	5.43	0.000	0.077	4.90	0.000
$GDP2$	-0.002	-3.91	0.000	-0.001	-2.95	0.003
HCE_{-10}	2.045	4.82	0.000	2.022	4.49	0.000
$HCE2_{-10}$	-0.565	-3.60	0.000	-0.491	-2.95	0.003
ALC_{-10}	-0.043	-0.70	0.484	0.022	0.23	0.822
$ALC2_{-10}$	0.002	1.02	0.307	-0.001	-0.22	0.827
CAL_{-10}	0.786	0.65	0.515	-0.723	-1.08	0.281
$CAL2_{-10}$	-0.083	-0.50	0.614	-0.135	1.53	0.125
$constant$	18.572	9.11	0.000	16.883	13.94	0.000
Wald $\chi^2(8)$:	210.64			338.33		
$Prob > \chi^2$:	0.0000			0.0000		
$1 - \frac{var(\varepsilon_{it})}{var(RLEF)}$:	0.601			0.559		
N	303			303		

[a]Note: Included in this Equation are (frequency): Australia(21), Austria(29), Belgium(22), Canada(12), Denmark(18), Finland(17), Germany(13), Ireland(15), Japan(2), Luxembourg(4), New Zealand(19), Norway(29), Portugal(15), Spain(13), Sweden(16), Switzerland(25), United Kingdom(18), U.S.A.(15)

Contemporaneous GDP (in US$) is associated with higher RLE in both sexes (with females possibly affected more strongly), confirming results found by Zweifel and Ferrari (1992) as well as Frech III and Miller (1999). A decreasing marginal effect of income cannot be excluded. The earlier study by Zweifel and Ferrari is also confirmed in that HCE_{-10} results in higher current RLE for both sexes. Since the squared explanatory variable proves highly significant, a critical value can be calculated beyond which additional lagged HCE is counter-productive (flat-of-the curve medicine). Among women, this is approximately US$ 1,800, among men, US$ 2,060, compared to a mean value of US$ 1,375 in 2000. With regard to the lifestyle indicators, (lagged) consumption of calories and alcohol do not show an influence. Similar findings would have been attained with a lag of 3 rather than 10 years.

Influence of *HCE* on 65+ Population Share

In order to determine the pressure of individuals faced with a given *RLE*, their population share must be accounted for, resulting in the variable *SISYPH* [defined in Equation (2.3)]. However, the population share of individuals 65 and older might depend on (lagged) HCE.

Table 2.3: Share of individuals aged 65 and older, 1970-2000

POP65[a]	Coefficient	z	$P > z$
GDP	0.115	6.64	0.000
$GDP2$	-0.002	-4.67	0.000
HCE_{-10}	0.626	1.35	0.177
$HCE2_{-10}$	-0.170	-0.97	0.333
ALC_{-10}	0.174	2.75	0.006
$ALC2_{-10}$	-0.005	-2.89	0.004
CAL_{-10}	-0.274	-1.40	0.161
$CAL2_{-10}$	0.050	1.80	0.071
constant	11.270	14.15	0.000
Wald $\chi^2(8)$:	332.63		
$Prob > \chi^2$:	0.0000		
$1 - \frac{var(\varepsilon_{it})}{var(POP65)}$:	0.190		
N	394		

[a]Note: Included in this Equation are (frequency): Australia(21), Austria(30), Belgium(30), Canada(30), Denmark(19), Finland(19), France(2), Germany(14), Ireland(30), Japan(2), Luxembourg(12), Netherlands(2), New Zealand(20), Norway(30), Portugal(20), Spain(29), Sweden(17), Switzerland(26), United Kingdom(25), U.S.A.(16)

The results of Table 2.3 do suggest endogeneity of *POP65*, but not with regard to HCE_{-10}. Specifically, higher *GDP* goes along with a higher share of aged people (up to some US$ 28,750), whereas *HCE* does not have a discernible effect. The first of these two statements holds also for a lag in *HCE*, *CAL* and *ALC* of 3 years; however, HCE_{-3} would have attained statistical significance. In the life style domain, alcohol consumption clearly serves to increase this share, but with decreasing effect at the margin. The other lifestyle variable does not contribute to the determination of *POP65*, regardless of the choice of lag.

Influences on Mortality

Since mortality is not a core variable of the Sisyphus syndrome but mainly serves to take the surge in *HCE* towards the end of human life into account, no gender-specific regressions are presented in Table 2.5. In accordance with early research on the United States [Auster et al.

(1969)], higher contemporaneous *GDP* might possibly increase mortality at the high end of the sample (beyond US$ 35,000; the same holds for HCE_{-10}, beyond US$ 1,490, driving home the argument made above that mortality occurs sooner or later). None of the lifestyle variables are statistically significant. All of these statements prove robust to a lag of 3 rather than 10 years.

Table 2.4: Influences on mortality, 1970-2000

$MORT^a$	Coefficient	z	$P > z$
GDP	-8.319	-3.46	0.001
$GDP2$	0.118	1.84	0.065
HCE_{-10}	-203.305	-5.68	0.000
$HCE2_{-10}$	68.349	4.02	0.000
ALC_{-10}	-16.843	-1.85	0.065
$ALC2_{-10}$	0.452	1.78	0.075
CAL_{-10}	86.145	1.32	0.188
$CAL2_{-10}$	-15.036	-1.78	0.075
$constant$	1025.007	8.42	0.000
Wald $\chi^2(8)$:	431.15		
$Prob > \chi^2$:	0.0000		
$1 - \frac{var(\varepsilon_{it})}{var(MORT)}$:	0.574		
N	351		

[a]Note: Included in this Equation are (frequency): Australia(19), Austria(30), Belgium(26), Canada(28), Denmark(16), Finland(16), Germany(13), Ireland(27), Luxembourg(10), New Zealand(19), Norway(28), Portugal(19), Spain(27), Sweden(14), Switzerland(21), United Kingdom(24), U.S.A.(14)

2.4.2 Explaining the Health Share in the GDP

In keeping with arguments proffered in section 2.2.1, the dependent variable here is *HCE/GDP*. In the earlier estimation of the Sisyphus syndrome by Zweifel and Ferrari (1992), an attempt was made to distinguish between private and public *HCE* to see whether the feedback from *RLE* to *HCE* is particularly marked in the public domain. However, the published literature does not make this distinction because the assignment of total *HCE* to these two components is far from perfect in many countries, resulting in measurement error (we thank Pedro Pito Barros, Lisbon, for calling our attention to this). With *HCE/GDP* being the dependent variable, resulting in a filtering out of common trends, measurement errors are expected to play a more important role to begin with. Therefore, the distinction between the private and public components is not made here. All endogenous variables estimated in section 2.4.1 enter with their predicted values, using a lag of 10 years (except for GDP).

Again, the fixed-effects (FE) and the random-effects (RE) specification is based on a Hausman test. This time, FE dominates. This divergence from the health production part of the model is puzzling at first sight. However, the production of health depends largely on internationally available medical technology, whereas HCE (as argued at the end of section 2.2.1), is also determined by institutional factors that are not so well presented exclusively by an error component whose autocorrelation coefficient is an average over all countries. Indeed, Canada and Germany differ comparatively little (some two percentage points) from the United States (the reference country), Australia, Austria, Portugal and Switzerland form an intermediate group (around three percentage points), while Luxembourg, Spain, and the United Kingdom are estimated to lie four percentage points below the United States, ceteris paribus. Other countries can not be assigned to one of the groups, either because they fall in between or because they had to be dropped due to multicollinearity (15 cases).

In principle, there could be a problem in particular with the lower bound being zero for $\frac{HCE}{GDP}$. However, calculations using extreme value combinations of explanatory variables inducing low predicted values of $\frac{HCE}{GDP}$ still produced estimates well within the positive range. This result is mainly due to the small effect of GDP on the health share. In spite of the political debate on its rise in major industrial countries, no significant effect of GDP can be discerned. This means that the estimated income elasticity of HCE is 1.14 (apart from the Sisyphus effect, see below).[2] This is lower than most elasticities in Gerdtham et al. (1992) but higher than the estimates presented in Gerdtham et al. (1998). As to the role of relative price RPH, the theoretical prediction would be that both private and public decision makers tend to substitute health goods for other commodities (aggregated in the GDP) when RPH is high. The positive coefficient of RPH indicates that the implied price elasticity is slightly below one.[3] Another prediction, already of relevance to the Sisyphus syndrome, is not confirmed,

[2]The income elasticity of HCE can be calculated as follows. Using $HCE \equiv (HCE/GDP)*GDP$, one obtains $\frac{\partial HCE}{\partial GDP} = \frac{\partial(HCE/GDP)}{\partial GDP} * GDP + \frac{HCE}{GDP}$. The first derivative equals $\beta_1' + \beta_3' \frac{\partial \widehat{MORT}}{\partial GDP} + \beta_4' \frac{\partial \widehat{SISYPH}}{\partial GDP}$ from Equation (2.4); therefore, the elasticity is given by $e(HCE, GDP) = (\beta_1' + \beta_3' \frac{\partial \widehat{MORT}}{\partial GDP} + \beta_4' \frac{\partial \widehat{SISYPH}}{\partial GDP}) \times \frac{(GDP)^2}{HCE} + 1 = 1.14$ evaluated at sample means (see Tables 2.1 and 2.5).

[3]Since HCE/GDP is formed using nominal figures, it contains RPH. Noting that $RPH = PH/PGDP$ (the ratio of the health to the GDP deflator), and using lower case symbols for real quantities, one has $(hce/gdp) * RPH = ...\beta_2' * RPH + ...$, or $hce * RPH = ...\beta_2' * RPH * gdp$. Differentiating with respect to RPH, one obtains, $hce * (1 + \eta) = \beta_2' * gdp$, where η is the elasticity of the demand for health care (hce) w.r.t. RPH.

Table 2.5: Health share in the GDP, 1970-2000

HCE/GDP	n[a]	Coefficient	z	$P > z$
$Dum\_Australia$	19	-3.057	-7.90	0.000
$Dum\_Austria$	26	-3.167	-8.62	0.000
Dum_Canada	12	-2.179	-5.35	0.000
$Dum\_Denmark$	16	-2.466	-6.26	0.000
$Dum\_Finland$	15	-3.238	-8.21	0.000
$Dum\_Germany$	11	-1.640	-3.89	0.000
$Dum\_Ireland$	15	-2.706	-6.27	0.000
$Dum\_Luxembourg$	4	-4.375	-8.45	0.000
$Dum\_NewZealand$	16	-3.679	-9.29	0.000
Dum_Norway	28	-3.515	-9.64	0.000
$Dum\_Portugal$	8	-2.890	-6.33	0.000
Dum_Spain	10	-4.001	-8.83	0.000
$Dum\_Switzerland$	21	-3.081	-8.14	0.000
$Dum\_UnitedKingdom$	17	-3.906	-9.89	0.000
GDP		-0.020	-1.13	0.260
RPH		0.012	2.03	0.042
$\widehat{MORT}$		0.010	1.72	0.085
$\widehat{SISYPH}$		0.069	2.98	0.003
$constant$		-18.092	-1.59	0.112
Wald $\chi^2(18)$:		524.26		
$Prob > \chi^2$:		0.0000		
$1 - \frac{var(\varepsilon_{it})}{var(HCE/GDP)}$:		0.847		
N		232		

[a]Note that n is the number of country-specific observations, with the benchmark U.S.A. being in the sample 14 times.

viz. that a higher mortality rate in a given year means a greater number of individuals in their last year of life, causing HCE to be high in that year. The coefficient of $\widehat{MORT}$ is positive but nonsignificant.

Of course, the variable of crucial interest is $\widehat{SISYPH}$, the remaining life expectancy at age 60 weighted by the share of individuals close to retirement age. It has a highly significant coefficient, possibly raising the scepter of an aging population that keeps the spiral in motion indefinitely.

Solving for η, one obtains $\eta = \beta_2' * (HCE/GDP)^{-1} - 1$. Evaluated at the sample means, this amounts to -0.99.

Conclusion 3. *The reestimation of the feedback relationship, linking (population share-weighted) remaining life expectancy to the health share in the GDP, yields preliminary statistical evidence of a Sisyphus syndrome.*

2.4.3 Assessing the strength of the Sisyphus syndrome

The estimates presented in the previous section allow to conclude that there is some preliminary evidence that between 1970 and 2000 a Sisyphus syndrome may have existed. The open question remains of whether the syndrome may be a phenomenon of sufficient importance to merit attention for some or all OECD countries. This calls for the evaluation of the partial derivative $\partial HCE/\partial HCE_{-10}$: If in a given year during the period of observation, HCE was stepped up, to what extent will later HCE be higher as a consequence of the syndrome (the lag being 10 years as established through specification tests)? If this derivative should exceed unity, this would constitute evidence of a locally explosive cycle; a globally explosive situation can be excluded because all derivatives to be evaluated have critical values where they become zero (linear and quadratic regressors have opposite signs in Tables 2.2 to 2.4). The identity

$$HCE \equiv (HCE/GDP) * GDP \tag{2.5}$$

may be used to obtain Equation (2.6), establishing the link between HCE/GDP and HCE_{-10}:

$$\begin{aligned}\frac{\partial HCE}{\partial HCE_{-10}} &= \\ &= \frac{\partial((HCE/GDP) * GDP)}{\partial HCE_{-10}} \\ &= \frac{\partial(HCE/GDP)}{\partial HCE_{-10}} * GDP + (HCE/GDP) * \frac{\partial GDP}{\partial HCE_{-10}} \\ &= \frac{\partial(HCE/GDP)}{\partial HCE_{-10}} * GDP \end{aligned} \tag{2.6}$$

since $\partial GDP/\partial HCE_{-10}$ is assumed to be zero due to the long lag.[4]

[4]However, other contributions found a significant influence of (contemporaneous) HCE on the GDP rate of growth [see e.g. Devarajan et al. (1996) or Beraldo et al. (2007)]. Estimating the elasticity of GDP with respect to HCE_{-10} at the sample means yields $e(GDP, HCE_{-10}) = 0.37$. This estimate is at the high end because it fails to control for increased inputs of labor and capital as well as technological change. In addition, replacing sample means by minimum and maximum values for the year 2000 yields $e(GDP, HCE_{-10})\ \epsilon$ (-0.681, 0.378), a range compatible with the assumption introduced.

Now HCE_{-10} influences current HCE through $SISYPH$, which is designed to capture the impact of population-weighted RLE on the demand for health care and hence HCE [see Equation (2.3) again]. Implicit differentiation then yields,

$$\begin{aligned}\frac{\partial(HCE/GDP)}{\partial HCE_{-10}} &= \\ &= \frac{\partial(HCE/GDP)}{\partial SISYPH} * \frac{\partial SISYPH}{\partial HCE_{-10}} \\ &= \frac{\partial(HCE/GDP)}{\partial SISYPH}\frac{1}{2}\left[\frac{\partial RLEF}{\partial HCE_{-10}} + \frac{\partial RLEM}{\partial HCE_{-10}}\right] * POP65, \end{aligned} \tag{2.7}$$

since $\partial POP65/\partial HCE_{-10}$ may be set to zero, according to Table 2.3. Substituting (2.7) into (2.6) and noting that GDP is measured in thousands of US$ and therefore needs to be rescaled, one obtains,

$$\begin{aligned}\frac{\partial HCE}{\partial HCE_{-10}} &= \\ &= \underbrace{\frac{1}{2}} * \underbrace{\frac{\partial(HCE/GDP)}{\partial SISYPH}} * \underbrace{\left[\frac{\partial RLEF}{\partial HCE_{-10}} + \frac{\partial RLEM}{\partial HCE_{-10}}\right]} * \underbrace{POP65} * \underbrace{GDP} * \underbrace{\frac{1}{1000}} \\ &= 0.5 * \quad 0.069 \quad * \quad (1.46 + 1.51) \quad * \; 13.60 \; * 15.76 * 0.001 \quad \approx 0.021\end{aligned} \tag{2.8}$$

The derivatives come from Tables 2.2 and 2.5 but have to be evaluated at specific values because the regressors of Table 2.2 appear not only in linear but also in quadratic form. Neglecting estimated coefficients that are nonsignificant and inserting overall sample means one concludes that US$ 1 of HCE_{-10} induces some US$ 0.02 of current HCE due to the dynamics of the Sisyphus syndrome. Assuming exponential decay, one could say that 68 percent of a one-time shock to HCE is carried over to the following year ($0.68^{10} \approx 0.02$). This is far enough from unity to suggest that the cycle is not explosive.

Conclusion 4. *The Sisyphus syndrome seems to have been of some importance during the observation period, with an estimated* 68 *percent of one US$ of additional HCE carrying over to the following year, which points to stability.*

2.5 Conclusion

This chapter started from the notion that the production of health and the determination of health care expenditure (*HCE*) should be analyzed jointly. Short of adopting the duality argument to the effect that estimating either a production function or an expenditure (share) function is sufficient, the two links are seen as constituting a dynamic feedback relationship. Once *HCE* is stepped up exogenously, it may well (with a lag) increase remaining life expectancy and hence the planning horizon of an individual considering an investment in health. This should increase demand for HCE, resulting in a feedback that constitutes the Sisyphus syndrome. The macroeconomic importance of this syndrome crucially depends on the remaining life expectancy of those who are at retirement age, beyond which they contribute little to the financing of HCE. These considerations call for the construction of the variable $S\widehat{ISYP}H$, which is the remaining life expectancy at age 60 weighted by the share of individuals aged 65+, the age discrepancy caused by data availability.

From 1970 to 2000 and among OECD countries, there is confirming evidence for both links. Prior HCE is positively related both to remaining life expectancy at age 60 and the population share of the aged 65+ and hence the $S\widehat{ISYP}H$ variable. In its turn, $S\widehat{ISYP}H$ proves to be a significant determinant of the health share in GDP. An initial HCE increase of US$ 1 is estimated to still result in US$ 0.02 a decade later as a result of the Sisyphus syndrome. Assuming exponential decay, 68 percent of a one-time shock thus are estimated to carry over to the following year. Since the income elasticity of *HCE* is estimated to be close to one ceteris paribus, much of the observed increase of the health share in the *GDP* can be traced to the Sisyphus syndrome.

However, one might argue that a Sisyphus syndrome does not really exist. For one, *GDP* could have been interpreted as a lifestyle variable rather than indicating a budget constraint. This would have called for the introduction of lags for this variable, which could have pointed to delayed adjustments of $\frac{HCE}{GDP}$ even exceeding the 30 years covered by the present sample. In addition, in view of the high degree of multicollinearity between *HCE* and *GDP* regardless of the lag selected (simple correlation coefficients always in excess of 0.8538), the estimated effects of lagged *HCE* on remaining life expectancy likely would have dropped and possibly could have lost significance. In that event, there would be no evidence of *HCE* contributing to the growth of a population segment that has a particular interest in the increased provision of health care services in a later period. Finally some OECD countries have missing values more often than others, which may introduce a selection bias to the extent that those not reporting have less problems with rising *HCE* (which are interpreted as a reflection of the Sisyphus syndrome in this chapter). The countries more consistently represented in the sample might therefore introduce a bias in favor of finding such a syndrome.

While such arguments cannot be rejected out of hand, there are at least two considerations that speak against these. The first is that the hypothesis of a Sisyphus syndrome as a dynamic relationship puts emphasis on lags. Specifically, in previous research on the production of health, contemporaneous *HCE* was not found to consistently lower mortality rates or increase longevity. However, when introducing a lag of ten years, both Zweifel and Ferrari (1992) and Frech III and Miller (1999) were able to relate remaining life expectancy to previous *HCE*, and this lag is replicated in the present study. Second, the link between production and factor demand theory suggests use of an expenditure share function (HCE/GDP) which is much less influenced by a time trend than *HCE* itself. And it is with such a detrended dependent variable that $\widehat{SISYPH}$, whose construction is entirely motivated by the maintained hypothesis, proves highly significant.

On the whole, the evidence suggests that a Sisyphus syndrome may have been operative in the OECD countries until the end of the past century. Although its dynamics are far from explosive, it mirrors the concern of policy makers that the decision to increase health care expenditure could have some similarity with the proverbial ride on a tiger.

Bibliography

Auster, R., I. Leveson, and D. Sarachek (1969). The production of health, an explanatory study. *The Journal of Human Resources 4*(4), 411–436.

Barros, P. P. (1998). The black box of health care expenditure growth determinants. *Health Economics 7*(6), 533–544.

Beraldo, S., D. Montolio, and G. Turati (2007). Healthy, educated and wealthy: Is the welfare state really harmful for growth? *Department d'Hisenda Publica. Universitat de Barcelona.*

Christensen, L. R., D. W. Jorgenson, and L. J. Lau (1973). Transcendental logarithmic production frontiers. *Review of Economics and Statistics 55*(1), 28–45.

Cochrane, A. L., A. S. St. Leger, and F. Moore (1978). Health service input and mortality output in developed countries. *Journal of Epidemiology and Community Health 32*(3), 200–205.

Devarajan, S., V. Swaroop, and H.-F. Zou (1996). The composition of public expenditure and economic growth. *Journal of Monetary Economics 37*, 313–344.

Frech III, H. E. and R. D. Miller, Jr. (1999). *The Productivity of Health Care and Pharmaceuticals: An International Comparison.* Washington D.C.: The AEI Press.

Gerdtham, U. G. and B. Jönsson (1991). Conversion factor instability in international comparisons of health care expenditure. *Journal of Health Economics 10*(2), 227–34.

Gerdtham, U. G. and B. Jönsson (2000). International comparison of health care expenditure. In A. J. Culyer and J. P. Newhouse (Eds.), *Handbook of Health Economics*, Volume 1A, Chapter 1, pp. 11–53. Amsterdam, London, New York, Oxford, Paris Shannon, Tokyo: Elsevier.

Gerdtham, U. G., B. Jönsson, M. MacFarlan, and H. Oxley (1998). The determinants of health expenditure in the OECD countries: A pooled data analysis. In P. Zweifel (Ed.), *Health, the Medical Profession, and Regulation. Developments in Health Economics and Public Policy, vol. 6*, pp. 113–34. Boston and Dordrecht: Kluwer Academic.

Gerdtham, U.-G. and C. J. Ruhm (2006). Deaths rise in good economic times: Evidence from the OECD. *Economics and Human Biology 4*, 298–316.

Gerdtham, U. G., J. Sögaard, B. Jönsson, and F. Andersson (1992). A pooled cross-section analysis of the health care expenditures of the OECD countries. In P. Zweifel and H. E. Frech III (Eds.), *Health Economics Worldwide. Developments in Health Economics and Public Policy Series, vol. 1, Norwell, Mass*, pp. 287–310. Boston and Dordrecht: Kluwer Academic.

Grossman, M. (2000). The human capital model. In A. J. Culyer and J. P. Newhouse (Eds.), *Handbook of Health Economics*, Volume 1A, Chapter 7, pp. 347–408. Amsterdam, London, New York, Oxford, Paris Shannon, Tokyo: Elsevier.

Hadley, J. (1982). *More Medical Care, Better Health? An Econometric Analysis of Mortality Rates*. Washington D.C.: Urban Institute Press.

Hadley, J. (1988). Medicare spending and mortality rates of the elderly. *Inquiry 25*(4), 485–493.

Kleiman, E. (1974). The determinants of national outlay on health. In M. Perlman (Ed.), *The Economics of Health and Medical Care*. London, Basingstoke: Macmillan.

Leu, R. E. (1986). The public-private mix and international health care costs. In A. J. Culyer and B. Jönsson (Eds.), *Public and Private Health Services: Complementarities and Conflicts*, pp. 41–63. Oxford and New York: Blackwell.

Lubitz, J., L. Cai, E. Kramarow, and H. Lentzner (2003). Health, life expectancy, and health care spending among the elderly. *The New England Journal of Medicine 349*(11), 1048–55.

Lubitz, J. D. and G. F. Riley (1993). Trends in medicare payments in the last year of life. *New England Journal of Medicine 328*, 1092–1096.

Maxwell, R. J. (1981). *Health and Wealth*. Lexington: Lexington Books.

Miller, Jr., R. D. and H. E. Frech III (2004). *Health Care Matters. Pharmaceuticals, Obesity, and the Quality of Life*. Washington D.C.: AEI Press.

Newhouse, J. P. (1977). Medical care expenditure: A cross-national survey. *Journal of Human Resources 12*(1), 115–125.

OECD (2001). *OECD Health Data*. Paris.

Stewart, Charles T., J. (1971). Allocations of resources to health. *The Journal of Human Resources 6*(1), 103–122.

Zweifel, P., S. Felder, and M. Meier (1999). Ageing of population and health care expenditure: A red herring? *Health Economics 8*(6), 485–496.

Zweifel, P. and M. Ferrari (1992). Is there a sisyphus syndrome in health care? In P. Zweifel and H. E. Frech III (Eds.), *Health Economics Worldwide. Developments in Health Economics and Public Policy Series, Vol. 1, Norwell, Mass*, pp. 311–30. Boston and Dordrecht: Kluwer Academic.

Life-Cycle Effects of Social Security in an Open Economy: A Theoretical and Empirical Survey

Peter Zweifel and Patrick Eugster*

Published in the "Zeitschrift für die Gesamte Versicherungswissenschaft", 97(1), 2008, 61-77.

*The authors would like to acknowledge the excellent research assistance by Michael Breuer as well as perceptive criticisms and suggestions by Robert Clark (North Carolina State University, Raleigh) and the participants of SUERF 2006 (Lisbon, Portugal).

Chapter 3

Life-cycle effects of social security in an open economy: A theoretical and empirical survey

3.1 Introduction

The conventional view of the relationship between demographic and social security (alternatively, the term social insurance will be used) is that exogenous demographic change has impacts both on the financing and the benefits side of social security. Usually, the economy is treated as a closed system, abstracting from both trade flows and migration of factors of production. If taken into account, these flows tend to be subsumed under 'effects of globalization', which often serve as a scapegoat for the future problems several social security systems are facing.

This chapter purports to propose a radically different view, which however is firmly rooted in economic theory. It recognizes the fact that demographic change is nothing but the aggregate of decisions made by individuals in the course of their life cycle. For example, the low share of young individuals in a population can ultimately be traced to the decision of potential parents not to have children or of actual parents to refrain from having additional children. As will be argued below, this (and other such decisions) may be influenced by social security. Therefore, the vantage point of analysis will be turned around, with emphasis placed on the impact social insurance has on individual decisions, which in the aggregate manifest themselves as demographic change.

The following methodology will be adopted. Throughout, the argument will be couched in terms of a marginal change (i.e. a moderate expansion or curtailment of social security) rather than in terms of two worlds, one with and the other without social security. While comparing two entirely different worlds would provide for sharp contrasts and crisp hypotheses, empirical evidence comes from the analysis of marginal changes of existing systems. In addition, the financing side is always distinguished from the benefits side. For example, an increase of contributions levied by a scheme for the provision for old age (OA) may occur without any adjustment of claims to future benefits, or benefits may be curtailed without a decrease in contributions. For example, in 2003 the Swiss federal government lowered the multiplier used for translating contributions into claims against the public pension system by decree in an attempt at staving off a deficit. Measures of this type cause the taxation element inherent in social security in the interest of redistribution to be enhanced, since current beneficiaries continue to obtain benefits in excess of the present value of their contributions [see Breuer (1999)].

In some systems (such as Germany and Switzerland), however, additional contributions paid do result in additional claims against the public OA scheme, at least in principle. For simplicity, this link will be neglected in the analysis, also because it does not exist in social health insurance (HI) and long-term care insurance (LCI), the two other branches of social security that presumably have a connection to decisions with demographic implications [other branches such as unemployment insurance and family allowances are neglected either because their connection to the decisions analyzed here is too weak or because they are less important financially, cf. OECD (1996)].

Finally, the analysis will follow an individual over a stylized life cycle characterized by an educational phase, marriage, and creation of a family, possible divorce, retirement, and death. This means that certain aspects of work life (such as the length of the work week and duration of unemployment) are disregarded. However, decisions concerning migration are taken into account where appropriate.

3.2 Educational effort and social insurance

3.2.1 Financing of social security and educational effort

This section expounds the impact increasing contributions to HI, OA and LCI may have on schooling. From an economic perspective, contributions to HI and LCI are roughly comparable to a tax on labor income. As always, such a tax has a substitution and an income effect. It makes leisure as the alternative to labor income less costly, causing it to be demanded in greater quantity, to the detriment of work (substitution effect). However, the reduction in disposable income occasioned by social security contributions may be important enough to cause individuals with a low labor income to work more rather than less (income effect). These two impacts on lifetime labor supply will be addressed again in a separate section. For the majority of workers in industrial countries, the income effect is unlikely to dominate, resulting in a positive wage elasticity of labor supply. Indeed, a recent German study incorporating life cycle effects finds a wage elasticity of labor supply of 0.6 for women and 0.8 for men [Fenge et al. (2002)]. Conversely, higher contributions to social security are thus predicted to discourage labor supply.

An expansion of social security therefore is expected to result in a permanent decrease of expected labor income over the life cycle. This has an impact on educational effort as soon as young decision makers consider education not only a cultural enrichment but also an investment. Expected future returns from additional educational effort therefore will be lower, causing some of possible investment on schooling not to occur. One would therefore expect an expansion of social security in the case of HI and LCI to have a negative effect on educational investment. This impact may be mitigated to some extent if vocational training (where usually contributions to social security must be paid) is substituted by university studies (where contributions usually are waived).

Table 3.1: Predicted effects of social security's financing on educational effort

	OA	HI	LCI
Effect on education	–	–	–

In the case of OA, the precise nature of the link between contributions and future benefits is decisive [see e.g. Lau and Poutvaara (2006)]. Köthenbürger and Poutvaara (2006) model a closed economy in which reducing social security contributions encourages investment in human capital. A similar result was obtained by Echevarria and Iza (2006), complemented by a simulation using different parameter values. However, contributions may give rise to claims more than proportionally (as is typical of buildup phases). During this time OA may even serve to increase the rate of return on educational investment and therefore induce additional educational effort. Indeed, such an effect was found by Flam and Risa (1995) in their empirical study. All effects are summarized in Table 3.1.

Turning to the role of increased globalization and economic openness, social security may induce migration of the labor force. As long as there are differences between countries in terms of contribution rates to OA (compared to benefits), workers may seek to earn their incomes in a country offering them a more favorable benefit-contribution ratio than their home country. Therefore, social insurance in the (less expensive) foreign country may achieve a surplus from migration, while social insurance in the (expensive) home country risks to end up in deficit. However, it was impossible to find empirical evidence of migration of workers induced by social security arbitrage.

3.2.2 Social security benefits and educational effort

With regard to HI, increased benefits should have an enhancing effect on educational effort. For HI secures access to medical care designed to maintain health status at a level that enables the young to pursue an education at all. Moreover, it reduces the risk of a future health loss that (due to lack of medical care) is so grave as to result in a partial or full loss of labor income, an effect that is of particular value to highly educated individuals.

In most industrial countries, OA benefits guarantee a minimum retirement income regardless of educational effort. In this way, OA causes a reduction of the marginal return to investing into education. However, this impact should be of minor importance at the time of deciding about education since rational individuals will discount benefits of OA heavily as they occur much later (see Table 3.2). In fact, there is some evidence that the subjective rate of discount is between 6 and 20 percent p.a. [Lawrance (1991); Coller and Williams (1999)]. On the other hand, OA may encourage the choice of occupations that yield a higher return to education due to increased risk [Anderberg (2000)]. With regard to globalization, Poutvaara (2007) constructs a theoretical model with mobile workers who decide about their investment in education. He shows that if the home country offers flat-rate OA benefits, then its workers invest more than in a closed economy because they hold the prospect of moving to the foreign country, where OA benefits increase with income.

Table 3.2: Predicted effects of social security's benefits on educational effort

	OA	HI	LCI
Effect on education	(−)	+	(−)

The benefits of LCI have two objectives. One is to prevent people from ending up in poverty for the last years of their life, the other, less prominent, to secure their bequeathable wealth. Concerning the first objective, LCI has the same effect as a supplement to OA and should therefore tend to somewhat discourage educational effort. As to the second objective, bequeathable wealth mainly stems from labor income earned over the life cycle. Protection of this wealth adds to its utility and hence should encourage educational effort as a means to create it. However, the predicted impact of LCI is minimal because final wealth will be positive only with a relatively low probability and because there must be relatives and friends worthy of the bequest. Finally, not only are these effects reduced by the respective probability weightings but also by discounting to present value (see Table 3.2). Not surprisingly, it has proved impossible to find supporting empirical evidence.

Conclusion 5. *The financing of the three branches of social security analyzed has a negative predicted overall impact on educational effort beyond the legal minimum. The ambiguous (OA, LCI) or countervailing (HI) influences of benefits are deemed to be too weak to counterbalance the clearly negative ones of contributions.*

The macroeconomic importance of this conclusion is evident. If social security should end up discouraging educational effort, it jeopardizes the future competitiveness of a country. With wage rates and employment increasing at a slower pace, this in turn creates problems for the financing of social security. This feedback effect constitutes a first indication that assuming economic growth to be exogenous to social security, as is usual for predicting its financial status, is deficient in that this disregards the effect social insurance itself may have on growth [see e.g. Echevarria and Iza (2006)].

3.3 Marriage and social insurance

3.3.1 Financing of social security and marriage

Often marriage provides the opportunity for one of the two partners to withdraw from the labor market. This decision is the more likely the smaller the concomitant loss of labor income. Since disposable labor income is reduced by social security contributions, nonmarket work becomes more attractive. In most countries, claims to HI and LCI are not affected as long as the working partner continues to pay contributions, while OA benefits are reduced to married couples in most OECD countries to reflect cost saving achieved by jointly living (Table 3.3 lists all effects). To the extent that mainly the less qualified worker shifts his or her activity into the household, increasing openness of the economy should reinforce this effect. The reason for this is the fact that globalization puts pressure especially on the wage rates of less qualified workers, thus reducing the opportunity cost of withdrawal [most likely women who on average continue to earn a lower wage rate than their male partners, cf. Disney and Whitehouse (2002)].

Table 3.3: Predicted effects of social security's financing on marriage

	OA	HI	LCI
Effect on marriage	(+)	+	+

Unfortunately, there does not seem to be empirical evidence regarding these predictions. The closest analogy is income taxation, which also serves to reduce disposable income (without however giving rise to future claims). The effects of taxation on marriage have been the subjects of empirical research. For example, Alm and Whittington (1997) find that the changing tax burden associated with marriage has an effect on the timing of marriage.

3.3.2 Social security benefits and marriage

Two persons living together in one household benefit from the advantage of risk diversification. Specifically, a minor health loss incurred by one partner can be taken care of by the other without having recourse to medical services. However, HI makes this advantage less decisive by facilitating access to formal medical services. HI could therefore cause marriage contracts not to be concluded or concluded later in life. However, in some countries (such as

Germany), HI covers nonactive family members as well without a surcharge. In these countries, this constitutes an advantage that may neutralize or even overcompensate the negative impact HI otherwise would have on marriage.
Similar arguments can be made with regard to OA. In some countries, married women are also covered without an additional contribution. This feature of OA benefits constitutes a definite incentive favoring marriage. Once more however, benefits are far in the future, causing them to be discounted to present value and reducing their impact.

Table 3.4: Predicted effects of social security's benefits on marriage

	OA	HI	LCI
Effect on marriage	(+)	–	–
	[0 also possible]	[(+) also possible]	

With regard to LCI, the notable fact is that the married spouse provides the major part of long-term care [Schneekloth and Pothoff (1993)]. With LCI, these informal services tend to be substituted by formal ones provided in homes and by professional service providers. Thus, LCI serve to somewhat undermine the motivation for marriage (its effects again occurring far in the future). Some preliminary evidence refers to the United States and a proposal to extend Medicaid to cover families with two rather than one parent present [Yelowitz (1998)]. Moreover, teenage marriages and pregnancies out of wedlock are found to be more frequent when public support (such as Aid for Mothers with Dependent Children, AFDC) is substantial in comparison to obtainable labor income [Bernstam and Swan (1986)]. These benefits would be lost if parents were to marry or live jointly in a household. Table 3.4 shows the considered effects of social insurance benefits on marriage. An opening of the economy probably would not have an additional impact, and there are no empirical studies investigating these additional effects.

Conclusion 6. *The financing side of social insurance creates incentives in favor of marriage, while the benefits have mixed predicted effects. OA benefits favor married couples, while HI and LCI tend to have weak negative effects on marriage.*

On the whole, social security may have an impact on the civilian status of a population. This impact likely has limited influence on economic development, with the exception of an increase of the aggregate cost of health care because of a further growth of the number of single-person households, who have to rely to a greater extent on the benefits of HI and LCI than married individuals [Zweifel and Breyer (1997), ch. 11.2].

3.4 Number of children and social insurance

3.4.1 Financing of social security and number of children

Additional contributions to be paid to social security (HI, OA, and LCI) reduce disposable income over a major part of the life cycle. One possibility of neutralizing this reduction is to keep the number of children low or to refrain from having children. A negative relationship between social insurance and the number of children (who originally were a source of security in the family) has been theoretically argued by Felderer (1992). Empirical evidence has been provided by Cigno and Rosati (1996), showing that in England, Germany, Italy and the United States expansions of OA coverage were linked to a reduction in birth rates. However, other branches of social insurance not considered here might encourage having children. For example, German unemployment insurance, the replacement rate increases from 65 to 75% of the previous wage when there are children [Hujer and Schneider (1989)]. Table 3.5 summarizes the effects of financing on children.

Table 3.5: Predicted effects of social security's financing on the number of children

	OA	HI	LCI
Effect on the number of children	–	–	–

Economic openness is predicted to reinforce this negative relationship because children constitute an impediment to mobility, making it more costly for parents to migrate in an attempt to avoid an unfavorable benefit-contribution ratio of social insurance. However, there do not seem to exist empirical studies concerning such a reinforcement effect.

3.4.2 Social security benefits and number of children

Again, the different branches of social insurance need to be distinguished. In principle, children increase the risk of incurring medical expenditures. By covering this cost to a great extent, HI mitigates this risk, thus encouraging larger families.

With regard to OA, there is potential substitution between support provided by children to their retired parents and the benefits of OA. An expansion of OA quite likely causes a reduction of support provided by children (Table 3.6). The evidence comes mainly from developing countries where children still importantly contribute to the livelihood of their parents [Lillard and Willis (1997); Swidler (1986)]. More recently, Ehrlich and Kim (2007) have analyzed a panel of 57 countries covering 33 years. They find that birth rates are negatively related to OA benefits, holding constant the gender ratio, net growth of marriages, female labor force participation, and the gender ratio of educational attainment among others.

While marriage partners provide most informal long-term care, children (especially daughters) are an important source as well. To the extent that social insurance provides financial support for formal services, substitution again is likely to be triggered, causing children to curtail their provision of services. However, this effect usually lies far in the future at the time of marriage, and the likelihood of LTC is still low [although increasing, see Strüwe (1996), Schmähl and Rothgang (1996)], causing them to be second-order.

Table 3.6: Predicted effects of social security's benefits on the number of children

	OA	HI	LCI
Effect on the number of children	(–)	+	(–)

With migration between open economies children who live abroad have reduced value to potential parents. Indeed, Schrieder and Knerr (2000) find evidence suggesting that once children have left their native community in Cameroon, the amount of money remitted to their parents is low unless they stand to receive a bequest. This means that parents' efforts at influencing or even controlling children's behavior become costly and often ineffective. This is even more true when children know that thanks to OA and LCI there is no threat of their parents' spending their last years of life in poverty. Moreover, LCI helps securing their inheritance at any rate. Given these difficulties, the incentive to have children at all is weakened even more [Lindbeck et al. (1999)].

Conclusion 7. *Both contributions and benefits of social insurance are likely to cause a substitution of children by social security benefits, thus reducing population growth.*

3.5 Divorce and social insurance

As a rule, the effects of social insurance on divorce should be the mirror image of those on marriage.

3.5.1 Financing of social security and divorce

When considering dissolving the marriage, the partner concerned frequently must take into account that she will have to return to work. The more contributions to social insurance reduce disposable income from work, the less attractive this alternative becomes. This implies that the financing side of social insurance may stabilize marriage [an effect which may be viewed critically, see e.g. Estes (2004)]. Table 3.7 summarizes the effects.

Table 3.7: Predicted effects of social security's financing on divorce

	OA	HI	LCI
Effect on divorce	−	−	−

3.5.2 Social security benefits and divorce

Having HI makes it easier for singles to cope with sickness because HI enables them to rely on formal medical care with little if any out-of-pocket payment. As to OA, pensions of married couples usually fall short of the double of the pension of a single person in most industrial countries, reflecting the cost savings achievable by jointly living in a household [Disney and Johnson (2001)]. The closer the equivalence of differences in pensions and costs of living, the more OA benefits are neutral with regard to divorce. It is the LCI benefit that lowers the cost of divorce. Since singles rely less on unpaid long-term care services provided by the marriage. Thus, this serves to reinforce the effects of HI, which has already been found to facilitate singly living (see Table 3.8).

Table 3.8: Predicted effects of social security's benefits on divorce

	OA	HI	LCI
Effect on divorce	(−)	+	+

Studies relate the increase in the number of single older women since 1950 in the United States to the expansion of OA benefits [Costa (1999)]. This result probably has little to do with the openness of the economy (as in the case of the marriage decision, see the pertinent section above).

Conclusion 8. *While contributions to social insurance create an incentive to avoid divorce, its benefits work in favor of divorce. The effects of the latter presumably outweigh the former.*

However, an increase of the divorce rate induced by social insurance should not have much of an impact on the economy because it also causes the rate of labor participation to increase. In fact, one-third of the increase of female participation rates in the United States during 1960 to 1980 can be explained by the increase of divorce rates; moreover, women in marriages at risk work more hours than do others [Johnson and Skinner (1986)]. This increase of labor supply could even enhance growth at the macroeconomic level, giving rise to a positive feedback on the financing of social insurance. On the other hand, singles are more likely to claim benefits especially of HI and LCI [Klein (1996)], so the overall impact of this feedback may well be unfavorable. In other words, since the so-called biological rate of return of pay-as-you-go social security importantly depends on the rate of growth of the active population [Aaron (1996)], its effect on divorce tends to jeopardize the financial equilibrium of social insurance in the long run.

3.6 Retirement and social insurance

3.6.1 Financing of social security and retirement

An important motive for deferring retirement is the prospect of earning extra labor income. However, to the extent that contributions to social insurance do not create well-defined individual claims (HI, LCI), they make deferral less attractive while encouraging earlier retirement. In the case of OA, additional contributions usually do found claims to future benefits. Where this is not the case, workers tend to retire early [Boldrin et al. (1999)]. This makes the expected impact of OA benefits on labor supply ambiguous [for the role of future claims in the context of unemployment insurance, see Atkinson and Micklewright (1991)], however, their expansion on the whole seems to have contributed to earlier retirement [Viebrok (1997)]. Table 3.9 summarizes the effects.

Table 3.9: Predicted effects of social security's financing on early retirement

	OA	HI	LCI
Effect on early retirement	0 [(+) also possible]	+	+

In the context of globalization, this effect is but reinforced because the portability of benefits increases the choice of locations for life in retirement. Being very recent, these tendencies have not yet been empirically studied.

3.6.2 Social security benefits and retirement

With regard to HI, the following effect is to be expected theoretically. Since utilization of medical care entails minimum out-of-pocket expense, other impediments become decisive. These impediments are the opportunity costs associated with travel and waiting time. Their importance has been documented e.g. by Leu and Doppmann (1986) [cited in Zweifel and Breyer (1997), ch. 4.4]. In retirement, these time costs lose relevance. By retiring, individuals benefit from lower opportunity costs. Conversely, an expansion of HI arguably encourages earlier retirement (Table 3.10), which permits the insured to have unfettered access to medical care.

The connection with OA is more direct as OA benefits serve to mitigate the loss of income that would go along with retirement. Moreover, early retirement almost without exception does not entail an actuarially fair reduction of benefits. This of course cannot but increase the incentive to retire from active life. In the case of Germany, Börsch-Supan and Schnabel (1997) as well as Siddiqui (1997) were able to show the expected effect; for the case of Switzerland, see Baldenweg-Bölle (1998), ch. 7; for the Netherlands, see Heyma (2004); for the U.S., see Burtless (1986).

Table 3.10: Predicted effects of social security's benefits on early retirement

	OA	HI	LCI
Effect on early retirement	+	+	−

Migration as part of increased openness presumably reinforces these effects in two ways. First, by spending time abroad, workers get to know different legal systems and lifestyles, which serves to widen their choice of locations for retirement. And among those additional locations, there may be one that makes early retirement particularly attractive. Second, provided the elderly are successful in having contribution rates increased rather than their pensions curtailed, high-wage earners are predicted to go elsewhere [Leers et al. (2003)].

This train of thought also establishes a connection with educational effort (see above). Increasingly, this effort does not occur in schools but during professional activity. The earlier expected retirement, the shorter the payback period to educational investment, and the less therefore the incentive to invest. This is the conclusion reached by Jensen et al. (2004), who then compare retirement benefits to old-age benefits (which are paid regardless of employment status), with old-age benefits clearly inducing longer learning. However, there does not seem to be empirical evidence relating to a reduction of continued training caused by earlier retirement.

Turning to LCI, the need for long-term care usually occurs at a time when the child (usually the daughter) is approaching retirement herself. With less LCI coverage, many daughters presumably would choose retirement in order to have time for providing support to their parents. With ample LCI coverage, they tend to continue work. Supporting empirical evidence has been found for the United States by Ettner (1995) as well as Wolf and Soldo (1994). Therefore, LCI may work against a tendency to quit the work force. However, this effect is limited to those still rare cases where at least one elder in fact becomes a long-term care patient. Therefore, on the whole the likely effect of social insurance is to encourage early retirement.

Conclusion 9. *The financing of social insurance makes early retirement attractive. The benefits of health insurance (HI) and old age provision (OA) have the same effect, whereas long-term care insurance (LCI) may weakly encourage deferred retirement. On the whole, the three branches of social insurance analyzed tend to reduce labor supply at higher ages.*

As noted already in the context of Conclusion 1, this constitutes another instance where the impact of social security on aggregate labor supply and hence income likely is unfavorable, creating the potential for negative feedback that on the long run may drive social insurance schemes into deficit.

3.7 Life expectancy and social insurance

3.7.1 Financing of social security and life expectancy

Admittedly individuals cannot choose their time of death. However, they certainly can undertake efforts that influence their remaining life expectancy by investing in prevention and medical care. In each individual case, early death may occur in spite of these efforts; however, at the level of entire populations, such effects are discernible [Frech and Miller (2001), Zweifel et al. (2005)]. Therefore, the argument proceeds as though everyone had influence on his or her time of death.

Table 3.11: Predicted effects of social security's financing on life expectancy

	OA	HI	LCI
Effect on life expectancy	0	0	0

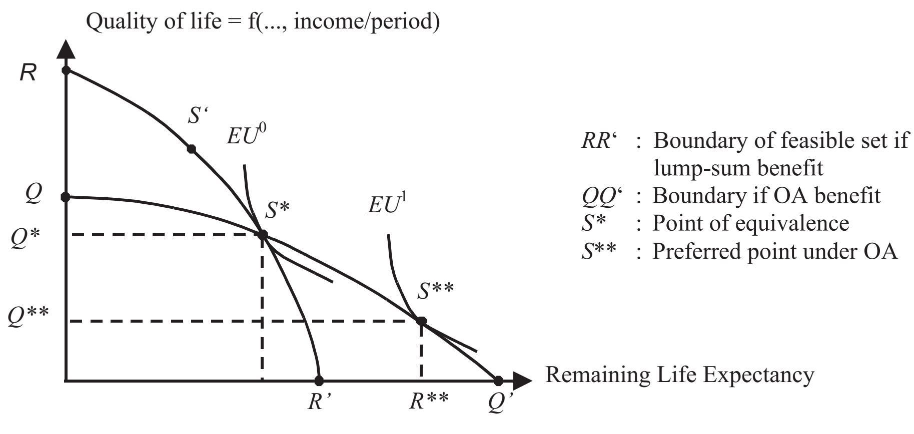
Since most individuals nowadays live to retirement, contributions to OA are a fixed cost once they are retired. As such, they do not have any relevance for efforts to increase life expectancy. As to the financing of HI and LCI, contributions do not increase with age (in spite of a growing risk that would call for an adjustment if they were risk-based as in commercial insurance). However, even private insurers fail to fully adjust HI premiums to this risk [Herring and Pauly (2001)]. This fact implies that as far as HI and LCI are concerned, the marginal cost of living another year is a constant; the financing side of social insurance thus is predicted to have no influence on efforts to increase remaining life expectancy (see Table 3.11), and there is no empirical evidence for such an influence.

3.7.2 Social security benefits and remaining life expectancy

Since OA benefits are paid as a regular income rather than a lump sum capital, longer life does not entail a reduced quality of life (due to fewer resources available per period). The contrast can be seen by comparing OA with a private life policy that pays a capital to survivors at a specified age. If beneficiaries had to live off this capital for the remainder of their lives, they would be constrained to spend less per period to balance an increase in longevity, causing a decrease in the quality of life. Longer remaining life expectancy would thus entail a financial sanction that is absent from current OA systems.

Figure 3.1 illustrates the argument. The curve RR' symbolizes the boundary of attainable combinations of remaining life expectancy and quality of remaining life years (determined at least in part by income per year) if a private insurer paid out a capital at age 60 (say). It indicates that more life years can only be had at the cost of a sacrifice in terms of quality of life. By way of contrast, the boundary QQ' symbolizes a typical public OA system, which guarantees an annual pension regardless of the number of years lived. As a matter of fact, the beneficiary needs to live another year to receive the OA benefit. On the other hand, preferences are represented by indifference curves that show a certain level of expected utility (EU). What Figure 3.1 shows is that while the individual seems to be indifferent between RR' and QQ' at point S^*, in fact he or she would opt for S^{**} on QQ' if given a choice, since $EU^1 > EU^0$. However, this choice goes along with a higher life expectancy R^{**} (compared to that pertaining to point S^*) and a lower quality of life. Also note that a transition to a point like S' (associated with reduced remaining life expectancy) is excluded if S^* was an initial optimum. Thus, the prediction is that current public pension systems encourage people to strive for a long life.

Figure 3.1: The predicted effects of a fixed-income vs. a lump-sum pension system



As to HI, it has the property of decreasing the price of medical care at the time of utilization, including those that serve to increase life expectancy. In the same vein, LCI subsidizes services that improve the quality of life, usually when time of death is near. The majority of beneficiaries have financial reserves that would last only for a short stay in a nursing home. Without LCI coverage, they would soon suffer a big reduction in quality of life that would undermine their interest in prolonging life. Therefore, the easing of the trade-off between longevity and quality of life noted in the case of OA is even more marked in the case of LCI.

Table 3.12: Predicted effects of social security's benefits on life expectancy

	OA	HI	LCI
Effect on life expectancy	+	+	+

The prediction that current OA systems increase remaining life expectancy was tested by Philipson and Becker (1998). Using information about retired employees of U.S. government, they distinguished groups according to their claims to future benefits. The group with the lowest expected benefits, amounting to less than US$ 1,000 a year, had 40 percent survivors at the age of 80 whereas the group with the highest, 63 percent. Clearly, it is inappropriate to attribute the entire difference to the incentives that emanate from the prospect of a higher OA benefit. Federal employees of the top group, being more educated and earning higher incomes, presumably would have had a higher remaining life expectancy even without the incentive of the OA benefit. Finally, these effects have little to do if at all with the degree of openness of an economy, mainly because they do not depend on the choice of residence in the country or abroad. Conclusion 6 summarizes Tables 3.11 and 3.12.

Conclusion 10. *Whereas contributions to social security (HI, OA, LCI) do not have an impact on longevity, clear incentives to prolong life emanate from benefits, which are conditioned on survival without exception. This effect is reinforced when the three branches of social security are considered in combination.*

Current institutional design not only of HI but also of the other branches of social security could therefore have at least some impact on the aging of population. Individuals profit considerably from not having to bear the risk of excessive remaining life expectancy. When deciding about contributions to social security, policy makers need to account for these effects lest they jeopardize the financial equilibrium of the system.

3.8 Conclusion

In this chapter, demographic change was not viewed as predetermined but endogenous, i.e. the consequence of individual decisions over the life cycle. The three branches of social security considered (health insurance HI, old age provision OA, and long-term care insurance LCI) have been found to influence individual decisions at several stages of the life cycle. The findings of this chapter are summarized in Table 3.13. The entries show that in the majority of cases, decisions are modified in a way as to lower labor supply and hence long-run economic growth, resulting in a reduction rather than increase of future contributions to social security. While many of the theoretically predicted impacts have not been systematically tested to this date, there is at least preliminary evidence supporting some of them. However, this evidence is not firm enough to permit estimation of a net effect where social security contributions and benefits have opposing impacts, e.g. divorce.

Table 3.13: Overview of theoretically predicted effects on life-cycle decisions

Decision	Effects of OA, HI, LCI					
	Contributions		Benefits		Globalization	
	Theory	Evid.	Theory	Evid.	Theory	Evid.
Education	– on wage	n.a.	+/– on wage	n.a.	+	n.a.
Marriage	– on LS	(✓)	+/– on LS	(✓)	0	n.a.
Children	+ on LS		(+) on LS	(✓)	+	n.a.
	– on future LS	✓				
Divorce	– on LS	n.a.	+ on LS	(✓)	0	n.a.
Retirement	– on LS	(✓)	– on LS	✓	+	n.a.
Life Exp.	0 on LS	(✓)	0 on LS	(✓)	0	n.a.
	RLE	(✓)	+ on RLE	(✓)		

Note: Evid. = evidence, Life Exp. = life expectancy, LS = labor supply, RLE = remaining life expectancy, (✓) = (some) confirming evidence, n.a. = not available

Clearly, the relationships expounded here are important also for private insurance. However, private insurance differs from social security in three aspects. First, private insurers make less favorable risks pay higher premiums, thus providing them with an incentive to improve their risk profile. Second, private insurers are very much aware of moral hazard, i.e. the changes in behavior induced by the very conclusion of an insurance contract. For example, many of them write policies in HI with cost sharing or with limited choice of physicians and hospitals (managed care options) in order to limit moral hazard effects. Third, the contracts written by private insurers do not have (at least under conditions of competition) the degree of homogeneity that characterizes social security, where homogeneity even is an objective in itself. Therefore, their incentive effects are not as bundled and unidirectional as those emanating from social security.

An important finding is that globalization, i.e. the transition to more open economies, tends to exacerbate the negative feedbacks an expansion of social security may have on economic growth. Systems run into problems sooner mainly because some of those who can benefit from increased mobility seek to avoid the financial burdens imposed on them by social insurance. In sum, the life-cycle effects of social security in open economies serve to constrain policy makers in their pursuit of the re-distributional goals they have traditionally been seeking to attain.

Bibliography

Aaron, H. (1996). The social insurance paradox. *Canadian Journal of Economics and Political Science 32*, 371–374.

Alm, J. and L. A. Whittington (1997). Income taxes and the timing of marital decisions. *Journal of Public Economics 64*(2), 219–240.

Anderberg, D. (2000). Occupational risks, social insurance and investment in education. *Journal of Population Economics 13*, 425–441.

Atkinson, A. B. and J. Micklewright (1991). Unemployment compensation and labor market transitions: A critical review. *Journal of Econonomic Literature 29*(Dec.), 1679–1727.

Baldenweg-Bölle, U. (1998). *Kollektive Altersvorsorge und flexibles Rentenalter in der Schweiz - (Collective provision for old age and flexible retirement in Switzerland).* Berlin: P. Lang.

Bernstam, M. and P. L. Swan (1986). The state as the marriage partner of last resort: New findings on minimum wage, youth joblessness, welfare, and single motherhood in the USA 1960-1980. *Stanford Hoover Institution Working Paper in Economics E-86-62.*

Boldrin, M., S. Jimenez-Martin, and F. Peracchi (1999). Social security and retirement in Spain. In J. Gruber and D. Wise (Eds.), *Social security and retirement around the world.* NBER.

Börsch-Supan, A. and R. Schnabel (1997). Social security and retirement in Germany. In J. Gruber and D. Wise (Eds.), *International Comparison of social security systems.* Chicago.

Breuer, M. (1999). Zahlungswiederstände bei Beitrags- versus Steuerfinanzierung in der Sozialversicherung: Wie gross sind die Unterschiede wirklich? - (resistance against paying for contribution vs. tax-financed social security: How important are the differences really?). In *Hamburger Jahrbuch für Wirtschafts- und Gesellschaftspolitik - (Hamburg's Yearbook of Political Economy and Political Society)*, pp. 45–67.

Burtless, G. (1986). Social security, unanticipated benefit increases, and the timing of retirement. *Review of Economic Studies 53*(5), 781–805.

Cigno, A. and F. C. Rosati (1996). Jointly determined saving and fertility behavior: Theory and estimates for Germany, Italy, UK and USA. *European Economic Review 40*(8), 1561–1589.

Coller, M. and M. B. Williams (1999). Eliciting individual discount rates. *Experimental Economics 2*, 107–127.

Costa, D.-L. (1999). A house of her own: Old age assistance and the living arrangements of older nonmarried women. *Journal of Public Economics 72*(1), 39–59.

Disney, R. and P. G. Johnson (2001). *Pension systems and retirement income across OECD countries.* Aldershot: Edward Elgar.

Disney, R. and E. Whitehouse (2002). The economics of well-being of older people in international perspective: A critical review. In S. Crystal and D. Shea (Eds.), *Annual Review of Gerontology and Geriatrics*, pp. 59–94.

Echevarria, C. A. and A. Iza (2006). Life expectancy, human capital, social security and growth. *Journal of Public Economics 90*(12), 2323–2349.

Ehrlich, I. and J. Kim (2007). Has social security influenced family formation and fertility. *Pharmaceutical Policy and Law 9*, 99–120.

Estes, C. L. (2004). Social security privatization and older women: A feminist political economy perspective. *Journal of Aging Studies 18*, 9–26.

Ettner, S. L. (1995). The impact of "parent care" on female labor supply decisions. *Demography 32*(1), 63–80.

Felderer, B. (1992). Does a public pension system reduce savings rates and birth rates? *Journal of Institutional and Theoretical Economics 148*, 312–325.

Fenge, P., S. Übelmesser, and M. Werding (2002). Second-best properties of implicit social security taxes: Theory and evidence. *CESifo Working Paper, No. 743*, Munich.

Flam, S. and A. E. Risa (1995). Market insurance, social insurance, and education. *Journal of Population Economics 8*(2), 149–160.

Frech, H. and R. D. Miller (2001). *The productivity of health care and pharmaceuticals: An international Comparison.* Washington D.C.: AEI Press.

Herring, B. and M. V. Pauly (2001). Premium variation in the individual health insurance market. *International Journal of Health Care Finance and Economics 1*(1), 43–58.

Heyma, A. (2004). A structural dynamic analysis of retirement behaviour in the Netherlands. *Journal of Applied Econometrics 19*, 739–759.

Hujer, R. and H. Schneider (1989). The analysis of labor market mobility using panel data. *European Economic Review 33*(2/3), 530–536.

Jensen, S. E. H., M. I. Lau, and P. Poutvaara (2004). Efficiency and equity aspects of alternative social security rules. *Finanzarchiv: Public Finance Analysis 60*(3), 325–358.

Johnson, W. R. and J. Skinner (1986). Labor supply and marital separation. *American Economic Review 76*(3), 455–469.

Klein, T. (1996). Determinants of institutionalization in old age. In R. Eisen and F. A. Sloan (Eds.), *Long-term care: Economic issues and policy solutions*, pp. 103–113. Boston.

Köthenbürger, M. and P. Poutvaara (2006). Social security reform and investment in education: Is there scope for pareto improvement? *Economica 73*(290), 299–319.

Lau, M. I. and P. Poutvaara (2006). Social security incentives and human capital investment. *Finish Economic Papers 19*(1), 16–24.

Lawrance, E. C. (1991). Poverty and the rate of time preference: Evidence from panel data. *Journal of Political Economy 99*(1), 54–77.

Leers, T., L. Meijdam, and H. A. A. Verbon (2003). Ageing, migration and endogenous public pensions. *Journal of Public Economics 88*(1-2), 131–159.

Leu, R. E. and R. J. Doppmann (1986). Die Nachfrage nach Gesundheit und Gesundheitsleistungen - (the demand for health and health care services). In G. Gäfgen (Ed.), *Ökonomie des Gesundheitswesens - (Economics of health care systems)*, pp. 161–175. Berlin.

Lillard, L. A. and R. J. Willis (1997). Motives for intergenerational transfers, evidence from Malaysia. *Demography 34*(1), 115–134.

Lindbeck, A., S. Nyberg, and J. W. Weibull (1999). Social norms and economic incentives in the welfare state. *Quarterly Journal of Economics 114*(1-35), 45–67.

OECD (1996). *Social expenditures statistic of OECD countries.* Paris: OECD.

Philipson, T. J. and G. S. Becker (1998). Old-age longevity and mortality-contingent claims. *Journal of Political Economy 160*(3), 551–573.

Poutvaara, P. (2007). Social security incentives, human capital investment and mobility of labor. *Journal of Public Economics 91*(7-8), 1299–1325.

Schmähl, W. and H. Rothgang (1996). The long-term costs of public long-term care insurance in Germany. Some guesstimates. In R. Eisen and F. A. Sloan (Eds.), *Long-term care: Economic issues and policy solutions*, pp. 181–222. Boston: Kluwer.

Schneekloth, U. and P. Pothoff (1993). Hilfs- und Pflegebedürftige in privaten Haushalten - (need for assistance and long-term care in private households). In R. Eisen and F. A. Sloan (Eds.), *Schriftenreihe des Bundesministeriums für Familie und Senioren - (Writings of the German ministry of family and senior citizens)*, pp. 20.2. Stuttgart: Kohlhammer.

Schrieder, G. and B. Knerr (2000). Labour migration as a social security mechanism for smaller households in Sub-Saharan Africa: The case of Cameroon. *Oxford Development Studies 28*(2), 223–236.

Siddiqui, S. (1997). Early retirement in West Germany: A sequential model of discrete choice. *Zeitschrift für Wirtschafts- und Sozialwissenschaften - (Journal of Economics and Social Sciences) 117*(3), 391–415.

Strüwe, W. (1996). Alterung und Pflegebedürftigkeit - (aging and need for long-term care). In P. Zweifel and S. Felder (Eds.), *Eine ökonomische Analyse des Alterungsprozesses - (An economic analysis of aging's process)*. Bern: Haupt.

Swidler, S. (1986). The old-age security motive for having children and the effect of social security on completed family size. *Quarterly Journal of Economics and Business 26*(2 Summer), 14–34.

Viebrok, H. (1997). *Das Arbeitsangebot im Übergang von der Beschäftigung in den Ruhestand - (Labor supply at the transition from employment to retirement)*. Frankfurt a.M.: Lang.

Wolf, D. A. and B. J. Soldo (1994). Married women's allocation of time to employment and care of elderly parents. *Journal of Human Resources 29*(4), 1259–1276.

Yelowitz, A. S. (1998). Will extending Medicaid to two-parent families encourage marriage? *Journal of Human Resources 33*(4), 833–865.

Zweifel, P. and F. Breyer (1997). *Health Economics*. New York: Oxford University Press.

Zweifel, P., L. Steinmann, and P. Eugster (2005). The sisyphus syndrome in health revisited. *International Journal of Health Care Finance and Economics 5*(2), 127–145.

Correlated Risks: A Conflict of Interest Between Consumers and Insurers and Its Resolution

PATRICK EUGSTER AND PETER ZWEIFEL*

Submitted to the "Journal of Risk Management in Financial Institutions".

*The authors would like to thank the participants of EGRIE 2006 (Barcelona, Spain) and SSES 2007 (St. Gallen) for valuable comments.

Chapter 4

Correlated risks: A conflict of interest between insurers and consumers and its resolution

4.1 Introduction

This chapter deals with a potential conflict between consumers and insurers which arises from their respective risk management objectives. Individuals as the potential purchasers of insurance need to manage three assets over their life cycle, health, wealth, and wisdom (skills of value in the labor market). Impulses impinging on these three assets tend to be positively correlated, serving to boost demand for insurance by risk-averse individuals. However, to the extent that benefits paid deviate from expected value (due to clauses in the small print on the one hand and customer accommodation policy on the other), these deviations should be negatively correlated across the different types of insurance held by an individual.

Insurers on their part shy away from covering risks that are positively correlated. While diversified shareholders might opt for the increase in total variance of underwriting results, they likely are concerned with the demand reaction of existing policyholders who dislike the increased risk of insolvency Cummins and Sommer (1996). Moreover, profitability may suffer because of the need for additional reserves to maintain solvency. Finally, management (having most of their wealth and wisdom tied to the company) seeks to limit underwriting risk.

This chapter purports to find out whether deviations of actual from expected insurance payments are indeed positively correlated across lines of insurance, which would serve to enhance demand but hamper supply, thus creating a conflict of interest between consumers and insurers. As in earlier work by Zweifel (2000) and Zweifel and Lehmann (2001), the investigation is carried out at the aggregate level. Deviations from expected value are associated with differences between observed and trend values of insurance payments. The countries studied are the United States and Switzerland for maximum contrast. Apart from the fact that the U.S. economy is 35 times bigger than the Swiss, there are important institutional differences. While private insurance in the United States has been rather competitive, a tight cartel imposed uniform products and premiums in Switzerland until 1992. In turn, Swiss social health insurance is individually contracted whereas U.S. employers (except small businesses) are mandated to offer health insurance policies as a fringe benefit to their workers. Compared to earlier research, much longer observation period (22 years or more), permitting improved tests for long-term trends are available this time. The absence of common trends in insurance payments would indicate a possibility for insurers to hedge their underwriting portfolio domestically or even across countries.

The next section (section 4.2) presents the relevant literature regarding demand for insurance in the presence of multiple risks and develops a test showing whether insurance payments reduce or increase volatility of personal assets. Section 4.3 describes the data, while section 4.4 explains the empirical approach. Section 4.5 shows the short-term and sections 4.6 and 4.7 the long-term results. Section 4.8 concludes.

4.2 Multiple risks: theory and stylized facts

4.2.1 Multiple risks in insurance economics

Multiple risks are of relevance for both the demand and supply of insurance. On the demand side, the classics by Arrow (1971) and Mossin (1968) show the conditions for demand for coverage to be positive if the relevant loss is associated with just one source of risk. Ideally, the contract would cover every possible loss regardless of its cause. However, both problems of asymmetric information (potential for both adverse selection and moral hazard) and regulation have prevented the writing of such contracts. The result is that markets for insurance are incomplete; see e.g. Doherty and Schlesinger (1983b), Mayers and Smith, Jr. (1983), Schlesinger and Doherty (1985), as well as Gollier and Schlesinger (1995) for an analysis of such markets.

Doherty and Schlesinger (1983b) and Eeckhoudt and Kimball (1992) show that independent insurance payments are suboptimal from the viewpoint of the consumer even in the case where losses are independent of each other. With two positively correlated losses, demand for insurance coverage is higher than in the case of independent losses, as demonstrated by Doherty and Schlesinger (1983a).

For insurers on the other hand, hedging their underwriting risks by combining lines of insurance that are negatively correlated seems natural. Positive correlation requires additional reserves for solvency and thus limits the supply of insurance. However, this argument overlooks three facts. First, assuming that insurers want to limit volatility, they can still hedge by holding assets that are positively correlated with their liabilities. After all, investment income importantly contributes to the bottom line of most companies. Second, as long as increased risk goes along with increased profitability, an insurance company may still be attractive to investors. The CAPM applied to insurance indicates that the underwriting beta is of crucial importance [Zweifel and Eisen (2003, ch. 6.2); Zweifel and Auckenthaler (2008)]. Finally, it is not clear that the owners of a company want management to eschew risk. Shareholders hold a call option whose value increases with the variance of the underlying, and it is the negative demand response of (informed) insurance purchasers that exerts a counter influence. However, Cummins and Sommer (1996) present empirical evidence suggesting that this counter effect is strong enough to cause insurers to limit their risk exposure. In sum, positive correlation of risks makes consumers increase their demand for insurance coverage but insurers to reduce their supply. These contradicting responses point to a conflict of interest.

4.2.2 How are assets of individuals correlated?

Individuals own three assets, health, wealth, and wisdom (skills generating labor income). They are subject to impulses that lead to variations in the total value of their portfolios.

Table 4.1 presents the seven impulses that are typically distinguished in (social) insurance, along with their hypothesized correlations. The impulses considered are illness, accident, disability, old age, unemployment, increase of family size, and death of the main breadwinner.

Table 4.1 can be read as follows. Given that someone is old (I4), there is probably a negative effect on increase in family size (I6). This points to a negative partial correlation. Family size in turn does not affect age; therefore, total correlation is negative. If on the other hand someone is sick at home in bed (I1), he or she is unlikely to cause an accident (I2), a negative partial correlation. However, it could also be that bad health is a result of an accident, a positive partial correlation. It is hypothesized that total correlation tends to be negative in this case since the first effect probably dominates.

Table 4.1: Impulses and their hypothesized correlations

	I1	**I2**	**I3**	**I4**	**I5**	**I6**	**I7**
I1: Illness							
I2: Accident	–?						
I3: Disability	−	+					
I4: Old age	+	+	+?				
I5: Unemployment	+	+?	+?	+?			
I6: Increase in family size	+?	0	0	−	+?		
I7: Death of main breadwinner	−	−	−	+	0?	0	

+ positive correlation; – negative correlation; 0 no correlation; +? possible positive correlation; –? possible negative correlation

Table 4.2 matches the seven impulses considered to the three assets affected. For instance, disability (I3) is assumed to have a strong influence on all three assets, while old age (I4) is hypothesized to influence health and wisdom but not wealth due to the fact that after retirement, wealth varies little with age (retired individuals consuming just about their pension).

Table 4.2: Impacts of impulses on individual assets

	Health	**Wealth**	**Wisdom**
I1: Illness	**	**	**
I2: Accident	**	**	**
I3: Disability	**	**	**
I4: Old age	*	0	*
I5: Unemployment	*	**	*
I6: Increase in family size	*	**	*
I7: Death of main breadwinner	*	**	0

** strong effect; * medium effect; 0 no effect

Combining the entries in Table 4.1 and Table 4.2, one sees that the three assets are likely to be positively correlated. Empirical research appears to support this notion. Bartel and Taubman

(1979), Mitchell (1990), and Ettner (1996) establish a positive correlation between health and wealth. Rosen and Taubman (1982) and Ashenfelter and Rouse (1998) find a positive relationship between wealth and wisdom, whereas Grossman (1975) and Kenkel (1991) present empirical evidence to the effect that wisdom and health are positively correlated.

Positive correlations between assets can lead to vicious cycles. Absent insurance, a person suffering a health loss will not only experience a drain on financial wealth but also have less opportunity to maintain the level of wisdom (through professional training). An optimal insurance system should avoid such vicious cycles by offsetting these losses. It is therefore important to determine how payments of insurance are correlated.

4.2.3 An application of portfolio theory

Since the seminal contribution of Markowitz (1959), it is accepted that risk-averse individuals seek to maximize expected return (μ) on their portfolio subject to a given amount of volatility (σ), i.e. attaining their efficient portfolio frontier in (μ, σ)-space. Now the different private and social insurance policies held by consumers can be viewed as assets constituting a portfolio. Since premiums are non-stochastic by and large, they will be ignored in the following. Payments, on the other hand contain a stochastic element for the insured. For example, ignored or forgotten clauses in the small print in private insurance and missing contribution years or change of employers in social insurance may cause payments to be below expected value. Conversely, consumer accommodation policy sometimes results in benefits that are higher than could be expected. In sum, there is reason for volatility in payments ($\sigma > 0$).

Table 4.3: Comparison of nominal rates of return on private and social insurance

	1960-2007	[0]	**1974-1979**	
	Private r_f [0]	Social l_w	Private r_f	Social l_w
United States	6.92	6.98	8.07	10.03
Switzerland	4.27	2.49[1]	4.90	*n.a.*
Sweden	8.22	7.69	9.62	11.87
Denmark	10.01	7.70[2]	16.38	13.17
United Kingdom	8.58	8.55	13.01	17.04
Austria	7.09[3]	6.23[4]	8.83	10.49
Japan	5.25[5]	8.33	7.90	13.46

Turning to expected returns (μ), there could considerably be differences between the different lines of insurance and particularly between private and social insurance. The rate of return on private insurance policies can be equated to the risk-free rate of interest on the capital market, while the one on social insurance can be approximated by the growth in wages, in keeping with the seminal argument by Samuelson (1958). However, no systematic difference between the two rates before tax can be inferred from Table 4.3. During the period 1960 - 2007, Switzerland, Sweden, Denmark, and Austria exhibit long-term rates of interest that exceed the growth of labor incomes, while the opposite holds true for Japan. However, during 1974 - 1979, the rate of interest was below the growth rate of wages and salaries both in Sweden and Austria. As an approximation, it seems legitimate to infer equality between the rates of return on private and social insurance. Additionally a conventional two-sided t-Test on the equality of means between the two time-series on return rates revealed a highly significant difference only for Denmark. In all the remaining countries the null Hypothesis on the equality of the two means could not be rejected on a one percent level. The (μ, σ)-criterion can therefore be reduced to volatility (σ). Accordingly, an ideal insurance system therefore should minimize total asset variance to which consumers are exposed.

Suppose there are two impulses, death of main breadwinner and accidents, affecting assets X and Y (e.g. health and wealth). Asset values in money terms, including expected payments for claims, are labelled X^a and Y^a (e.g. life insurance and accident insurance). Unexpected deviations from expected benefits are symbolized by x and y, respectively. Total asset variance then amounts to

$$\mathrm{Var}(X^a + x + Y^a + y) = \begin{cases} \underbrace{\mathrm{Var}(X^a)}_{+} + \underbrace{\mathrm{Var}(x)}_{+} + \underbrace{\mathrm{Var}(Y^a)}_{+} + \underbrace{\mathrm{Var}(y)}_{+} + \\ \underbrace{2 \cdot \mathrm{Cov}(X^a, Y^a)}_{+} + \underbrace{2 \cdot \mathrm{Cov}(X^a, x)}_{0} + \underbrace{2 \cdot \mathrm{Cov}(X^a, y)}_{0} + \\ \underbrace{2 \cdot \mathrm{Cov}(Y^a, x)}_{0} + \underbrace{2 \cdot \mathrm{Cov}(Y^a, y)}_{0} + \underbrace{2 \cdot \mathrm{Cov}(x, y)}_{?} \end{cases} \tag{4.1}$$

Four out of the six covariances can be set to zero for the following reasons. In the case of $\mathrm{Cov}(X^a, x)$ and $\mathrm{Cov}(Y^a, y)$, this is due to the very definition of unexpected deviations (x, y). In addition, $\mathrm{Cov}(X^a, y)$ and $\mathrm{Cov}(Y^a, x)$ are (close to) zero because normally insurance covers just one type of loss, regardless of the occurrence of another impulse. For example, premature death usually does not trigger a payment under the title of accident insurance. As to $\text{Cov}(X^a, Y^a)$, it is positive as long as the policies feature cost-sharing provisions because the two impulses still result in lowered asset values. In all, the sign of $\text{Cov}(x, y)$ proves crucial for total asset variance. If this covariance is negative, it reduces the total amount of asset variance, if positive, increases it. Thus, to the extent that correlations between unexpected deviations in insurance payments are positively correlated, they expose consumers to excess asset volatility, preventing them from reaching the efficient frontier in (μ, σ)-space.

In sum, positive correlations between stochastic deviations in insurance payments cause inefficiencies in the insurance system as a whole. This conclusion forms the basis of the empirical tests to be carried out in sections 4.4 to 4.7 below. Admittedly, correlation reflects linear dependence only. Recently, copula theory has been developed to model hidden nonlinear dependencies that might be responsible for fat tails in the loss distribution [see e.g. Embrechts et al. (2003), Embrechts (2008)].[6]

4.3 Data

The analysis deals with yearly data for insurance claims paid in the United States and Switzerland. They cover the period 1970 - 2004 in the case of the United States for private insurance and from 1980 - 2001 for social insurance. The data is taken from the *Life Insurers Fact Book 2003* and *2005*, published by the American Council of Life Insurers. They are complemented by information from *National Health Expenditures 2003 and 2004* published by the U.S. Department of Health & Human Services for the years 1960 - 2004. The data on social insurance come from the OECD Social Expenditure Database 2004, covering 1980 to 2001.

In the case of Switzerland, the source is *Die privaten Versicherungseinrichtungen in der Schweiz*, published by the Federal Agency for Private Insurance (Bundesamt für Privatversicherung) for the years 1975 - 2004. Data on benefits paid by social insurance again come from the OECD Social Expenditure Database 2004, covering the years 1980 - 2001. Explanations and labels for lines of insurance distinguished are given at the end of appendix B.

The information on GDP was taken from the national accounts statistics' published by the OECD, who revised the pre- 1989 figures for Switzerland. The data for the Swiss GDP is estimated from 1989 backwards. Comparison with the data provided by the Swiss Statistical

[6] However, most copulas deal with losses at the level of an individual company, whereas the present investigation uses data aggregated over all insurers of a country. In addition, the data entering calculations are deviations from trend values taken from an ordinary least squares estimation.

Office shows them to be almost identical. To keep matters simple, all data was taken from one source; they are in millions of national currency throughout.

4.4 Econometric specification

The most simple way to obtain time series estimates of unexpected short-term deviations from expected values is to define them as deviations from trend. Aggregate insurance payments increase progressively over time due to inflation. The estimation procedure therefore includes a linear and a quadratic time trend to accept for inflation. The same regression Equation is used for all insurance lines,

$$I_{i,t} = \beta_{i,1} + \beta_{i,2} \cdot T_t + \beta_{i,3} \cdot T_t^2 + \epsilon_{i,t}, \tag{4.2}$$

where I is payment in insurance line i, t denotes the year of observation, T_t is the time trend, and $\epsilon_{i,t}$ is the error term.

The first objective is to identify short-term correlations between regression residuals since individuals as well as insurers try to avoid positive correlations. In Zweifel (2000) as well as Zweifel and Lehmann (2001), positive correlations between payments prevail for the period 1975 - 1993.

A regression system for private, social, and both private and social insurance is estimated. This amounts to three regression systems for each country. Since numbers of observations differ, longer time series are reduced to the length of the shortest one. For the case of U.S. private insurance, the system reads,

$$\begin{pmatrix} PLID_t \\ PLIDI_t \\ PLAI_t \\ PHI_t \\ PHI2_t \end{pmatrix} = \begin{pmatrix} \beta_{1,PLID} & \beta_{2,PLID} & \beta_{3,PLID} \\ \beta_{1,PLIDI} & \beta_{2,PLIDI} & \beta_{3,PLIDI} \\ \beta_{1,PAI} & \beta_{2,PAI} & \beta_{3,PAI} \\ \beta_{1,PHI} & \beta_{2,PHI} & \beta_{3,PHI} \\ \beta_{1,PHI2} & \beta_{2,PHI2} & \beta_{3,PHI2} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ T_t \\ T_t^2 \end{pmatrix} + \begin{pmatrix} \epsilon_{PLID,t} \\ \epsilon_{PLIDI,t} \\ \epsilon_{PAI,t} \\ \epsilon_{PHI,t} \\ \epsilon_{PHI2,t} \end{pmatrix}, \tag{4.3}$$

with $E(\epsilon\epsilon') = \Omega$. A Breusch-Pagan chi-square test indicates that Ω cannot be considered diagonal; therefore, the error terms are correlated across Equations.

Next, each Equation is estimated individually for each line of insurance in order to maximize the number of observations. The calculation of pairwise correlations between residuals indicates whether unexpected negative (positive) deviations from trend in one line occur along with negative (positive) deviations from trend in another line of insurance within a given year. This is interpreted as positive short-term correlation at the level of the individual insured; a discussion of possible errors of aggregation taken from Zweifel (2000) is given in appendix A.

To analyze the common long-term trends that may characterize insurance payments, a cointegration analysis using the GDP as the cointegration variable is performed Greene (2003, ch. 20.4.3). First, time series need to be tested for their stationarity. Usually, one uses first differences; however, the objective here is to identify common trends rather than avoiding spurious correlation in specifying a model explaining the development of claims paid. Therefore, the variables are left in levels. Next, since two time series can only be cointegrated if they are of the same order, the integration order of each series is checked and reported. Finally, an augmented Engle-Granger test that accepts the null hypothesis of no cointegration will be sufficient to conclude that the two variables have no common trend. Critical values are taken from Davidson and MacKinnon (1993, ch. 20.6). If on the other hand the null hypothesis is rejected, then cointegration may be inferred even though strictly speaking it has not been verified Greene (2003, p. 656). So whenever the residuals of Equation (4) below exhibit stationary properties, a common trend can be said to exist:

$$I_{i,t} = \gamma_1 + \gamma_2 \cdot GDP_t + \mu_{i,t}, \tag{4.4}$$

with $\mu_{i,t}$ assumed i.i.d. This choice can be justified for two reasons. First, with income elasticities above one and rate elasticities low for major lines of insurance [SwissReinsurance (1989)], I_t is likely to be dominated by movements in GDP_t. Second, testing for cointegration with GDP is far easier than using one, several, or all remaining lines of insurance as variables of cointegration (however see section 4.7 below). The residuals serve as input to the augmented test based on a specification without a trend variable. The auxiliary regression then reads (dropping the i subscript for simplicity),

$$(\hat{\mu}_t - \hat{\mu}_{t-1}) = \varphi_2 \cdot \hat{\mu}_{t-1} + \varphi_3 \cdot (\hat{\mu}_{t-1} - \hat{\mu}_{t-2}) + \nu_t, \tag{4.5}$$

with ν_t assumed i.i.d. The parameter φ_2 provides an indication of whether I_t and GDP_t show some form of common trend. If this should be the case for several lines of insurance, then their payments would tend to move parallel in the longer term, permanently subjecting the insured to extra volatility of their assets. In addition, insurers would have no opportunity to combine the lines affected in order to hedge their underwriting risk within the country considered.

However, increasingly insurance companies operate internationally, which provides them with opportunities for global diversification. Therefore, as a last step, an additional cointegration analysis involving the private lines of U.S. and Swiss insurance is performed in section 4.7. If unexpected deviations of payments show a common trend across countries, then global diversification within the same line of insurance is possible. Since social insurance will continue to be national schemes for some time to come, this last step applies to private insurance in the United States and Switzerland only.

4.5 Analysis of short-term correlations

The short-term analysis tries to ascertain whether deviations from expected value of claims paid are positively or negatively correlated across different lines of insurance within the same country [see Equation (4.1)]. Section 4.5.1 deals with the United States, while section 4.5.2 is devoted to Switzerland.

4.5.1 United States

The evidence for the United States derived from the trend estimations are presented in Tables 4.4 to 4.6. In private insurance (Table 4.4), four out of ten correlations are significantly negative and just two are positive. Interestingly, it is payments of private health insurance as reported by the U.S. Department of Health & Human Services ($PHI2$), which cover the years 1960 - 2004 [rather than 1970 - 2004 as in the case of (PHI)] that correlates negatively with three of the other private insurance lines, viz. private life insurer's death payments ($PLID$), health insurance written by life insurers (PHI), and their annuity payments ($PLAI$). Disability payments ($PLIDI$) and annuity payments ($PLAI$) provided by life insurers also correlate negatively with each other.

Table 4.4: Correlations of trend deviations in U.S. private insurance, 1965 - 2004

	$PLID$	$PLIDI$	$PLAI$	PHI	$PHI2$
$PLID$	1.000				
$PLIDI$	−0.0534 (0.7435)	1.000			
$PLAI$	−0.0227 (0.8896)	−0.3417* (0.0282)	1.000		
PHI^a	0.4830** (0.0033)	0.1358 (0.4366)	0.4584** (0.0056)	1.000	
$PHI2^b$	−0.3514* (0.0262)	0.2105 (0.1922)	−0.4383** (0.0047)	−0.5516** (0.0006)	1.000

*,** coefficient significant at the five and one percent level, respectively
Note: The time series PHI and PHI2 refer to the same type of payments; only the source of data differs. Therefore, the negative correlation between the two is puzzling.
[a] 1970 - 2004
[b] 1960 - 2004

On the other hand, deviations in payments for health insurance by private life insurers (PHI) correlate positively with those in death ($PLID$) and annuity ($PLAI$) payments. This indicates a part of insurance business where consumers would benefit from risk diversification. One way could be to offer them combined policies; however, insurers appear to be unwilling to write such policies. On the whole, however, U.S. purchasers of private insurance are exposed to little excessive risk.

In U.S. social insurance, the picture is quite different, as evidenced in Table 4.5. Social disability benefits ($SDCB$), worker's compensation ($SWCB$), survivor's (SSB) and health (SHB) benefits show five significant positive correlations. Old age ($SOACB$), paid sick leave ($SPSB$), family cash ($SFCB$) and unemployment (SUB) benefits all show at least one significant positive correlation. In total 15 (or 54 percent) out of 28 correlations are significantly positive, the coefficients ranging from 0.44 to 0.94. Just two correlations indicate a negative relationship. U.S. social insurance thus seems to contribute little to asset stability for consumers. This result is in keeping with findings reported in Zweifel (2000), whose data base did not include OECD sources.

Table 4.5: Correlations of trend deviations in U.S. social insurance, 1980 - 2004

	$SDCB$	$SWCB$	$SOACB$	$SPSB$	SSB	$SFCB$	SUB	SHB
$SDCB$	1.000							
$SWCB$	0.9336**	1.000						
	(0.0000)							
$SOACB$	0.3033	0.0989	1.000					
	(0.1700)	(0.6615)						
$SPSB$	0.3115	0.3846	0.3477	1.000				
	(0.1582)	(0.0771)	(0.1128)					
SSB	0.7867**	0.6613**	0.4395*	−0.1444	1.000			
	(0.0000)	(0.0008)	(0.0407)	(0.5215)				
$SFCB$	0.7769**	0.7101**	0.2728	−0.0626	0.8888**	1.000		
	(0.0000)	(0.0002)	(0.2193)	(0.7821)	(0.0000)			
SUB	0.6085**	0.6291**	0.4030	0.5833**	0.3508	0.2444	1.000	
	(0.0027)	(0.0017)	(0.0629)	(0.0044)	(0.1094)	(0.2730)		
SHB	0.9443**	0.8588**	0.3194	0.1141	0.8884**	0.8863**	0.4690*	1.000
	(0.0000)	(0.0000)	(0.1473)	(0.6132)	(0.0000)	(0.0000)	(0.0277)	

*,** coefficient significant at the five and one percent level, respectively

Table 4.6 shows the cross-correlations between private and social insurance. The only line of private insurance showing significant (positive) correlation is health insurance ($PHI2$) as reported by the U.S. Department of Health and Human Services. Therefore, while U.S. private insurance fails to consistently make up for unexpected shortfalls in the benefits of its social counterpart (and vice versa), it at least does not cause much excess asset volatility for consumers.

In sum, U.S. private insurance seems to considerably reduce the volatility of personal assets, contrary to social insurance which may well expose consumers to excess volatility. The interplay between private and social insurance does not alleviate the problem, leaving asset volatility unaffected.

Table 4.6: Correlations of trend deviations in U.S. private and social insurance, 1980 - 2004

	$PLID$	$PLIDI$	$PLAI$	PHI	$PHI2$
$SDCB$	0.1436	0.2790	−0.1358	0.2016	0.5015*
	(0.5239)	(0.2086)	(0.5467)	(0.3682)	(0.0174)
$SWCB$	−0.0211	0.3357	−0.1055	0.1521	0.5314*
	(0.9256)	(0.1266)	(0.6403)	(0.4993)	(0.0109)
$SOACB$	−0.0268	0.0604	−0.2457	−0.2593	0.2701
	(0.9058)	(0.7895)	(0.2704)	(0.2440)	(0.2240)
$SPSB$	−0.3367	−0.1628	−0.3591	−0.4135	0.4683*
	(0.1255)	(0.4690)	(0.1007)	(0.0558)	(0.0280)
SSB	0.2115	0.3337	0.0304	0.3112	0.2641
	(0.3447)	(0.1290)	(0.8931)	(0.1586)	(0.2349)
$SFCB$	0.1113	0.3175	0.0619	0.3054	0.2840
	(0.6218)	(0.1500)	(0.7845)	(0.1670)	(0.2003)
SUB	−0.3437	0.0431	−0.3343	−0.3070	0.5746**
	(0.1173)	(0.8488)	(0.1284)	(0.1647)	(0.0052)
SHB	0.2407	0.3455	−0.1405	0.3005	0.3582
	(0.2806)	(0.1152)	(0.5330)	(0.1743)	(0.1017)

*,** coefficient significant at the five and one percent level, respectively

4.5.2 Switzerland

For Switzerland, the evidence is reported in Tables 4.7 to 4.9. In private insurance, there are three (out of six) significant correlations, all positive and ranging from 0.42 to 0.72 (Table 4.7). Two of them are due to life insurance ($PLID$), one to general liability insurance (PGI). Since all the remaining correlations suggest a positive relationship as well, it must be said that payments by Swiss private insurance tend to increase rather than reduce the volatility of consumers' assets. Again, this finding is in accord with those of Zweifel (2000), although the time series used here cover ten more years.

Table 4.7: Correlations of trend deviations in Swiss private insurance, 1975 - 2004

	$PLID$	PGI	PAI	PHI
$PLID$	1.000			
PGI	0.5262**	1.000		
	(0.0028)			
PAI	0.1507	0.4153*	1.000	
	(0.4266)	(0.0225)		
PHI	0.7223**	0.3252	0.1763	1.000
	(0.0000)	(0.0795)	(0.3514)	

*,** coefficient significant at the five and one percent level, respectively

Social insurance in Switzerland again shows a less favorable picture than in the United States. No less than 16 out of 36 correlations are significantly positive in Table 4.8, with a maximum value as high as 0.97. The main source of the problem is unemployment insurance (SUB), whose benefits exhibit a significant relationship with six other insurance lines. This reinforces the findings of Zweifel and Lehmann (2001), who report 6 positive out of 10 correlations for the period 1970 to 1990. Therefore, Swiss social insurance seems to cause consumers to be rather far away from their efficient frontier in (μ, σ)-space.

Table 4.9 shows the cross-correlations between Swiss private and social insurance. Whereas only one correlation coefficient is significantly negative, 6 out of 36 are clearly positive, with a maximum of 0.75. The responsibility for this lies mainly with the benefits of private accident insurance (PAI), which accounts for five positive relationships. Similar results were found by Zweifel and Lehmann (2001); they reinforce the impression that the United States insurance system does a much better job than its Swiss counterpart when it comes to filling unexpected shortfalls in the benefits paid by either private or social insurance.

Table 4.8: Correlations of trend deviations in Swiss social insurance, 1980 - 2004

	SDCB	*SOIB*	*SOACB*	*SPSB*	*SSB*	*SFCB*	*SHTB*	*SUB*	*SHB*
SDCB	1.000								
SOIB	0.1408	1.000							
	(0.6624)								
SOACB	−0.1789	0.2374	1.000						
	(0.4257)	(0.4576)							
SPSB	0.1936	−0.1700	−0.0032	1.000					
	(0.5465)	(0.5973)	(0.9921)						
SSB	−0.3090	0.0542	0.9725**	0.0035	1.000				
	(0.1617)	(0.8670)	(0.0000)	(0.9913)					
SFCB	0.4735*	0.1712	0.5982**	−0.1026	0.4949*	1.000			
	(0.0260)	(0.5947)	(0.0033)	(0.7510)	(0.0192)				
SHTB	−0.2508	0.2524	0.9443**	0.0692	0.9304**	0.5912**	1.000		
	(0.2602)	(0.4288)	(0.0000)	(0.8309)	(0.0000)	(0.0038)			
SUB	0.4595*	0.3374	0.6886**	0.0035	0.5797**	0.9617**	0.6652**	1.000	
	(0.0314)	(0.2835)	(0.0004)	(0.9914)	(0.0047)	(0.0000)	(0.0007)		
SHB	−0.2371	0.4868	0.7714**	0.2653	0.7536**	0.3740	0.7299**	0.5109*	1.0000
	(0.2881)	(0.1085)	(0.0000)	(0.4047)	(0.0001)	(0.0864)	(0.0001)	(0.0151)	

*,** coefficient significant at the five and one percent level, respectively

In all, the insurance system of Switzerland offers vast opportunity for efficiency improvements both in its private and its social components. Regarding the interplay between the two, the insurance system of both Switzerland and the United States fail to fill the gaps in one component by extra payments in the other.

Table 4.9: Correlations of trend deviations in Swiss private and social insurance, 1980 - 2004

	$PLID$	PGI	PAI	PHI
$SDCB$	−0.2741	−0.3367	−0.1105	−0.3242
	(0.2171)	(0.1255)	(0.6244)	(0.1411)
$SOIB$	0.4159	0.2214	0.4621	0.4739
	(0.1788)	(0.4892)	(0.1304)	(0.1196)
$SOACB$	−0.0075	0.3053	0.6116**	−0.1704
	(0.9735)	(0.1671)	(0.0025)	(0.4482)
$SPSB$	−0.2479	0.3025	0.2886	−0.0850
	(0.4372)	(0.3392)	(0.3630)	(0.7928)
SSB	−0.0279	0.3226	0.5258*	−0.1856
	(0.9018)	(0.1431)	(0.0120)	(0.4082)
$SFCB$	−0.3094	−0.1032	0.2932	−0.5406**
	(0.1612)	(0.6477)	(0.1854)	(0.0094)
$SHTB$	−0.0529	0.3570	0.5803**	−0.2590
	(0.8151)	(0.1029)	(0.0046)	(0.2445)
SUB	−0.2500	−0.0132	0.4251*	−0.4167
	(0.2619)	(0.9535)	(0.0486)	(0.0537)
SHB	0.1230	0.5396**	0.7491**	0.1940
	(0.5855)	(0.0096)	(0.0001)	(0.3869)

*,** coefficient significant at the five and one percent level, respectively

4.6 Long-term analysis

Especially for the insurer, year-to-year movements of claims are of limited relevance. When looking for hedging opportunities in their underwriting business, they need to know whether claims tend to move parallel in the longer term. This issue calls for cointegration analysis.

In cointegration analysis, the time series need to be of the same order of integration. A standard Dickey-Fuller (DF) test applied to first differences without a time trend suggests that most of the series are I(1) [see Table 4.10; the critical values are based on Fuller (1976) after interpolation]. If this test rejects the null hypothesis of a unit root, then this is reported in Table 4.10 as I(1). If the null hypothesis is not rejected, pointing to a higher order of

Table 4.10: Order of integration as indicated by the standard Dickey-Fuller test

	USA		**Switzerland**	
	t	I(1)?	t	I(1)?
$PLID$	−3.859***	I(1)	−3.425**	I(1)
$PLIDI$	−10.683***	I(1)		
$PLAI$	−6.034***	I(1)		
PHI	−4.714***	I(1)	−2.847*	I(1)
$PHI2$	−6.699***	I(1)		
PGI			−6.737***	I(1)
PAI			−4.485***	I(1)
$SDCB$	−1.480	not I(1)	−3.877***	I(1)
$SWCB$	−1.991	not I(1)		
$SOACB$	−3.557*	I(1),trend	−4.049***	I(1)
$SPSB$	−2.974*	I(1)	−3.411**	I(1)
SSB	−2.756*	I(1)	−4.148***	I(1)
$SFCB$	−4.109***	I(1)	−2.665*	I(1)
SUB	−4.087***	I(1)	−2.768*	I(1)
SHB	0.033	not I(1)	−1.816	not I(1)
$SOIB$			−2.890*	I(1)
$SHTB$			−4.723***	I(1)
GDP	−4.035**	I(1),trend	−3.019**	I(1)

*,**,*** coefficient significant at 10, 5, 1 percent level

integration, then an additional DF test with a time trend is performed. In all, three U.S. and only one Swiss time series cannot be said to be integrated of order one, i.e. I(1).

Additionally, the more powerful Dickey-Fuller GLS test (accounting for autocorrelation in the error term) as proposed by Elliot et al. (1996) is carried out. The results are presented in the appendix B; they largely coincide with those of Table 4.10. Whenever the conclusion of both tests is that a series is not I(1), it is excluded from the analysis below. If just one test indicates the differenced series to have a unit root, then it is still analyzed but the results appear in bold italics in Table 4.11 to indicate that they need to be interpreted with caution. This testing strategy results in the exclusion or qualification of U.S. life insurance ($PLID$) as well as workers' compensation ($SWCB$), health insurance (SHB), disability cash benefits ($SDCB$), and paid sick leave benefits ($SPSB$) in U.S. social insurance. In the case of Switzerland, social health insurance (SHB) had to be dropped.

As can easily be seen from Table 4.11, none of the remaining lines of private insurance are cointegrated with the GDP of the respective country. This means that there is an absence of

Table 4.11: Results of cointegration (cointegration variable: GDP)

	USA, 1975 - 2004		Switzerland, 1975 - 2004	
	φ_2	Cointegrated?	φ_2	Cointegrated?
$PLID$	−0.2409	***No***	−0.1441	No
$PLIDI$	−0.4669	No		
$PLAI$	−0.2128	No		
PHI	−0.3648	No	−0.1063	No
$PHI2$	−0.3092	No		
PGI			−0.7146[a]	No
PAI			−0.5706	No
$SDCB$	−0.3024*	***Yes***	−0.0963	No
$SWCB$				
$SOACB$	−0.6347*	Yes	−0.1532	No
$SPSB$	−0.2626	***No***	−0.7492**	Yes
SSB	−0.2970	No	−0.3212	No
$SFCB$	−0.2712	No	−0.2777	No
SUB	−0.6139	No	−0.4035	No
SHB	−0.3082	***No***		
$SOIB$			−0.7511	No
$SHTB$			−0.2793	No

*,**,*** coefficient significant at 5, 2.5, 1 percent level
[a]significant at the 10 percent level

common trends, creating hedging possibilities. Specifically, private U.S. health insurers might consider writing umbrella-type policies covering several risks. Turning to social insurance, one notes that possibly disability benefits ($SDCB$) and definitely old age benefits ($SOACB$) exhibit a common trend with GDP. The scope for hedging thus is more limited. In the case of Switzerland, private insurers could also create combined products that would serve to reduce asset volatility for consumers. The same is true of social insurance, with the sole exception of paid sick leave ($SPSB$), whose benefits are clearly trending with the Swiss GDP.

Finally, the question needs to be addressed of whether certain lines of private insurance are more amenable to hedging than others. This calls for performing a cointegration analysis with one of the other lines of insurance rather than GDP serving as the cointegration variable. In Tables 4.12 and 4.13, only domestic alternatives are considered. In the United States (Table 4.12), not a single line of business is cointegrated with any other, with the one possible exception of health insurance combined with life insurers' death payments. In Switzerland

(Table 4.13), accident insurance (PAI) and general liability insurance (PGI) share a significant common trend; otherwise there is no trace of parallel long-term movements.

Table 4.12: Cointegration of private insurance in the U.S.A.

	$PLID$		$PLIDI$	
	φ_2	Cointegrated?	φ_2	Cointegrated?
$PLID$				
$PLIDI$	−0.2330	No		
$PLAI$	−0.2796	No	−0.0980	No
PHI	−0.5080[a]	No	−0.1378	No
$PHI2$	−0.0691	No	−0.4320	No
	$PLAI$		PHI	
	φ_2	Cointegrated?	φ_2	Cointegrated?
$PLID$				
$PLIDI$				
$PLAI$				
PHI	−0.3604	No		
$PHI2$	−0.2353	No	−0.4397	No

[a]coefficient significant at the 10 percent level
*,**,*** coefficient significant at 5, 2.5, 1 percent level

Table 4.13: Cointegration of private insurance in Switzerland

	$PLID$		PHI	
	φ_2	Cointegrated?	φ_2	Cointegrated?
$PLID$				
PHI	−0.5001	No		
PGI	−0.1457	No	−0.1770	No
PAI	−0.1499	No	−0.1224	No
	PGI			
	φ_2	Cointegrated?		
$PLID$				
PHI				
PGI				
PAI	−0.8913*	Yes		

*,**,*** is significant at 5, 2.5, 1 percent level

In conclusion, payments of private insurance are predominantly devoid of common trends both in the United States and Switzerland, providing ample opportunity to offer combined insurance contracts. In social insurance, a few common trends can be identified which limit somewhat the scope for umbrella-type policies.

4.7 Possibilities of international diversification

Increasingly, insurance companies operate globally. Therefore, it is natural to ask whether claims development in the United States and Switzerland permit a degree of global hedging (which of course must be of limited extent in view of the much smaller Swiss market). Again, cointegration tests serve to provide evidence on this issue; however, contrary to section 4.6, they involve the same line of business in the United States and Switzerland. If the two time series turn out not to be cointegrated, there exists the possibility of international hedging for private multinational insurers. For example, they could offer life insurance both in the United States and Switzerland, benefiting from the fact that the two claims processes exhibit different trends. This choice is not available to social insurers for some time to come, motivating this investigation to be performed for private insurance only.

Throughout, the U.S. variable is used as the regressor. As shown in Table 4.14, all but one line of business (payments for death by Swiss life insurance $PLID-CH$, with U.S. life insurance annuity payments $PLAI-US$) are not cointegrated. These two lines of insurance are the only ones that should not be used for diversifying the companies' underwriting risk.

In the case of the United States and Switzerland, international diversification of the underwriting risk by private insurers therefore is entirely possible. These hedging opportunities go some way towards mitigating the conflict of interest between consumers (who would like to transfer positively correlated risks) and insurers (who eschew these risks) of the same country.

4.8 Conclusion

The starting point of this chapter is the argument that insurance payments contain a stochastic element. To the extent that these elements are positively correlated, they expose consumers to excess asset volatility. At the same time, insurers shy away from underwriting positively correlated risks. Equating these stochastic elements to trend deviations in aggregate insurance payments, the contribution of insurance to the reduction of consumers' asset volatility is reflected by the degree of negative correlation between the deviations in the differ-

Table 4.14: Cointegration tests of Swiss and U.S. lines of insurance

	$PLID-US$		$PLIDI-US$	
	φ_2	Cointegrated?	φ_2	Cointegrated?
$PLID-CH$	−0.2520	No	−0.2349	No
$PHI-CH$	−0.1986	No	−0.2213	No
$PGI-CH$	−0.3960	No	−0.2371	No
$PAI-CH$	−0.1750	No	−0.2155	No
	$PLAI-US$		$PHI-US$	
	φ_2	Cointegrated?	φ_2	Cointegrated?
$PLID-CH$	−0.3268*	Yes	−0.3683	No
$PHI-CH$	−0.1813	No	−0.2278	No
$PGI-CH$	−0.5823[a]	No	−0.4479	No
$PAI-CH$	−0.3738	No	−0.2853	No
	$PHI2-US$			
	φ_2	Cointegrated?		
$PLID-CH$	−0.4906	No		
$PHI-CH$	−0.2903	No		
$PGI-CH$	−0.5736	No		
$PAI-CH$	−0.2051	No		

[a]significant at the 10 percent level
*,**,*** is significant at 5, 2.5, 1 percent level

ent lines of private and social insurance. However, at least in the case of the United States and Switzerland, these deviations are found to often correlate positively, especially in Swiss social insurance. To overcome this inefficiency, insurers in the two countries could write combined policies. However, for this to be feasible in the long term, the claims processes must not exhibit common trends. Cointegration tests show this to be the case almost without exception. Finally, multinational insurers have the additional option of hedging a line of their business in the United States by underwriting the same type of risk in another country (Switzerland for example). This again proves to be possible with very few exceptions.

In sum, this research suggests that consumers both in the United States and Switzerland fail to reach their efficient frontiers in (μ, σ)-space as far as their assets 'insurance policies' are concerned. It also finds scope for additional hedging by private insurers, domestically as well as globally. However, some open questions need to be addressed in future research. First, similar investigations should be carried out for additional countries to check whether truly global hedging opportunities exist in the underwriting of personal insurance. Second, analysis of individual data would be extremely valuable (see the remarks in section 4.2.3 on copula theory as well). As already discussed in Zweifel (2000), aggregation creates the danger of finding positive correlations in the aggregate that do not exist at the individual level. Likewise, the results of the cointegration tests reported here need not carry over to individual insurance companies. However, the results presented in this chapter do seem to raise issues of considerable importance worthy of additional research effort.

Appendix A

This appendix is devoted to the criticism of testing an individual theory with aggregate data and is adopted from Zweifel (2000). Suppose two individuals i and j with assets X and Y experience deviations from their expected payments. Aggregate covariance is then given by,

$$\begin{aligned} \mathrm{Cov}(x_i + x_j, y_i + y_j) &= E(x_i + x_j)(y_i + y_j) \\ &= Cov(x_i, y_i) + \mathrm{Cov}(x_j, y_j) + \mathrm{Cov}(x_i, y_j) + \mathrm{Cov}(x_j, y_i). \end{aligned}$$

This formula is used to discuss three points:

1. *Positive correlation induced by macroeconomic factors* While within individual covariances could well be negative or zero the last two terms can nevertheless be positive. This could be due to a recession or some other macroeconomic shock. Suppose a recession triggers increased losses with regard to skills and health. If both payments must be financed out of a given budget for social insurance, then a budget cut might compel negative deviations in payment x_i to individual i and y_j to individual j yielding $Cov(x_i, y_j) > 0$, while within individual covariances remain zero. This is possible but unlikely.

2. *Numbers effect of macroeconomic factors* Since it is possible that within individual covariances are positive for all claimants, macroeconomic factors could increase total covariance by each additional individual considered. This however is not quite convincing. Aggregate deviations in unemployment insurance (leading proxy of macro influence) is

correlated with deviations in private insurance in 2 out of 9 cases (see Tables 4.6 and 4.9). Macroeconomic shocks do not seem to support the findings presented here.

3. *Failure to filter out nonexpected components* The estimation procedure uses a quadratic time trend. Residuals are assumed to be orthogonal in this estimation procedure. However, insurance benefits do not need to follow a quadratic time trend perfectly. In this case some parts of the estimated unexpected benefits are in fact expected which may be positively correlated whenever there is a common triggering event. This may be the case e.g. in the United States where social health benefits and social disability cash benefits correlate with $\rho = 0.94$. However, the correlation between social health benefits and private health insurance is not significantly positive in either country while they certainly share a common triggering event. The positive correlation of the expected payment did not spill over in unexpected deviations. So this cannot be decisive in every case.

The short discussion illustrates that no criticism provides an alternative consistent explanation of the observed phenomenon. Nevertheless the theory should be estimated using individual panel data to correctly answer all the questions addressed.

Appendix B

Table 4.15: GLS cointegration: optimal lags and order of integration

	United States			Switzerland		
	opt.lag	TS[a]	I(?)	opt.lag	TS[a]	I(?)
PLID	9	−1.840	not I(1)	3	−4.425***	I(1),trend
PLIDI	1	−5.382***	I(1),mean		−−	−
PLAI	5	−2.917*	I(1),trend		−−	−
PHI	0	−5.307***	I(1)[b],trend	5	−2.629*	I(1),trend
*PHI*2	0	−9.074***	I(1)[b],trend		−−	−
PGI		−−	−	1	−5.384***	I(1),trend
PAI		−−	−	1	−2.383*	I(1),mean
SDCB	4	−3.231**	I(1),trend	0	−4.976***	I(1)[b],trend
SWCB	0	−1.991	not I(1)		−−	−
SOACB	0	−3.557*	I(1)[b],trend	0	−4.074**	I(1)[b],trend
SPSB	6	2.400	not I(1)	0	−3.411**	I(1)[b],mean
SSB	7	−4.786***	I(1),trend	0	−4.083**	I(1)[b],mean
SFCB	0	−4.134**	I(1)[b],trend	0	−2.665*	I(1)[b],mean
SUB	7	−4.483***	I(1),trend	0	−2.768*	I(1)[b],mean
SHB	3	−3.378**	I(1),trend	0	−2.195	not I(1)
SOIB		−−	−	0	−3.554*	I(1)[b],trend
SHTB		−−	−	0	−4.704***	I(1)[b],trend
GDP	0	−4.035**	I(1)[b],trend	0	−3.091**	I(1)[b],mean

*,**,*** coefficient significant at 10, 5, 1 percent level

[a]TS = test-statistic

[b]Zero lag length indicated by Ng-Perron sequential t (1995); standard Dickey-Fuller test with interpolated critical values according to Fuller (1976)

Table 4.16: Labels for insurance benefits, with sources

United States	(P= private, S=social)
$PLID$	Life insurer's death payments; Life Insurers Fact Book (LIFB)
$PLIDI$	Life insurer's disability payments and retained assets; LIFB
$PLAI$	Life insurer's annuity payments; LIFB
PHI	Health insurance; LIFB
$PHI2$	Health insurance; U.S. Department of Health and Human Services
$SDCB$	Disability cash benefits; OECD Social Expendit. Database (OECD SED)
$SWCB$	Worker's compensation cash benefits; OECD SED
$SOACB$	Old age cash benefits; OECD SED
$SPSB$	Paid sick leave benefits; OECD SED
SSB	Survivor's benefits total; OECD SED
$SFCB$	Family cash benefits; OECD SED
SUB	Unemployment benefits; OECD SED
SHB	Health benefits; OECD SED
Switzerland	(P= private, S=social)
$PLID$	Life insurance; Bundesamt für Privatversicherungen
PGI	General liability insurance; Bundesamt für Privatversicherungen
PAI	Accident insurance; Bundesamt für Privatversicherungen
PHI	Health insurance; Bundesamt für Privatversicherungen
$SDCB$	Disability cash benefits; OECD SED
$SOIB$	Accident insurance (disability and survivors); OECD SED
$SOACB$	Old age cash benefits; OECD SED
$SPSB$	Paid sick leave benefits; OECD SED
SSB	Survivor's benefits total; OECD SED
$SFCB$	Family cash benefits; OECD SED
$SHTB$	Housing benefits total; OECD SED
SUB	Unemployment benefits; OECD SED
SHB	Health benefits; OECD SED

Bibliography

Arrow, K. J. (1971). *Essays in Theory of Risk Bearing.* Chicago: Markham Publishing Company.

Ashenfelter, O. and C. Rouse (1998). Income, schooling, and ability: Evidence from a new sample of identical twins. *The Quarterly Journal of Economics 113*(1), 253–284.

Bartel, A. and P. Taubman (1979). Health and labor market success: The role of various diseases. *The Review of Economics and Statistics 61*(1), 1–8.

Cummins, J. and D. Sommer (1996). Capital and risk in property-liability insurance markets. *Journal of Banking and Finance 20*, 1069–1092.

Davidson, R. and J. G. MacKinnon (1993). *Estimation and Inference in Econometrics.* New York: Oxford University Press.

Doherty, N. A. and H. Schlesinger (1983a). The optimal deductible for an insurance policy when initial wealth is random. *Journal of Business 56*(4), 555–565.

Doherty, N. A. and H. Schlesinger (1983b). Optimal insurance in incomplete markets. *Journal of Political Economy 91*(6), 1045–1054.

Eeckhoudt, L. and M. Kimball (1992). Background risk, prudence and the demand for insurance. In G. Dionne (Ed.), *Contributions to Insurance Economics*, pp. 239–254. Boston: Kluwer Academic Publishers.

Elliot, G., T. J. Rothenberg, and J. H. Stock (1996). Efficient tests for an autoregressive unit root. *Econometrica 64*(4), 813–836.

Embrechts, P. (2008). Copulas: A personal view. *Department of Mathematics, ETH Zurich, Switzerland*.

Embrechts, P., F. Lindskog, and A. McNeil (2003). Modelling dependence with copulas and applications to risk management. In S. Rachev (Ed.), *Handbook of Heavy Tailed Distributions in Finance*, pp. 329–384. Amsterdam: Elsevier.

Ettner, S. L. (1996). New evidence on the relationship between income and health. *Journal of Health Economics 15*(1), 67–85.

Fuller, W. A. (1976). *Introduction to Statistical Time Series.* New York: Wiley.

Gollier, C. and H. Schlesinger (1995). Second-best insurance contract design in an incomplete market. *Scandinavian Joournal of Economics 97*(1), 123–135.

Greene, W. H. (2003). *Econometric Analysis - Fifth Edition.* New Jersey: Pearson Education Internationl.

Grossman, M. (1975). The correlation between health and schooling. In N. E. Terleckyj (Ed.), *Household Production and Consumption.* New York: Columbia University Press.

Kenkel, D. S. (1991). Health behavior, health knowledge, and schooling. *Journal of Business 99*(2), 287–305.

Markowitz, H. M. (1959). *Portfolio Selection - Efficient Diversification of Investments.* New York: Wiley.

Mayers, D. and C. W. Smith, Jr. (1983). The interdependence of individual portfolio decisions and the demand for insurance. *Journal of Political Economy 91*(2), 304–311.

Mitchell, J. M. (1990). The effect of chronic disease on work behavior over the lifecycle. *Southern Economic Journal 56*(4), 928–942.

Mossin, J. (1968). Aspects of rational insurance purchasing. *Journal of Political Economy 76*(3), 533–568.

Rosen, S. and P. Taubman (1982). Changes in life-cycle earnings: What do social security data show? *Journal of Human Resources 17*(3), 321–338.

Samuelson, P. A. (1958). An exact consumption-loan model of interest with or without the contrivance of money. *Journal of Political Economy 66*(6), 467–482.

Schlesinger, H. and N. A. Doherty (1985). Incomplete markets for insurance: An overview. *Journal of Risk and Insurance 52*(3), 402–423.

SwissReinsurance (1989). Volkswirtschaftliche Bestimmungsfaktoren des Versicherungsgeschäfts - (Macroeconomic determinants of the development of insurance activity). *Sigma 7*.

Zweifel, P. (2000). Criteria for the future division of labor between private and social health insurance. *International Journal of Health Care Finance and Economics 26*(3), 38–55.

Zweifel, P. and C. Auckenthaler (2008). On the feasibility of insurers' investment policies. *Journal of Risk and Insurance 75*(1), 193–206.

Zweifel, P. and R. Eisen (2003). *Versicherungsökonomie - (Insurance Economics),* 2^{nd} *ed.* Berlin: Springer.

Zweifel, P. and H. Lehmann (2001). Soziale Absicherung im Portfolio persönlicher Aktiva: Wie kann die soziale Sicherung Teufelskreise verhindern? - (Hedging the portfolio of personal assets: How can social insurance avoid vicious cycles?). In E. Theurl (Ed.), *Der Sozialstaat an der Jahrtausendwende*, pp. 53–75. Heidelberg: Physica Verlag.

Capping Risk Adjustment?

Peter Zweifel, Patrick Eugster and Michèle Sennhauser*

Submitted to the "Journal of Health Economics".

*The authors would like to thank PD Dr. Konstantin Beck (University of Zurich and CSS insurance Lucerne) for his comments as well as the three health insurers that provided data for this study.

Chapter 5

Capping risk adjustment?

5.1 Introduction

When premiums are community-rated, risk adjustment is introduced in order to reduce competitive insurers' incentive to select favorable risks [for a general discussion on risk adjustment see Ellis and Van de Ven (2000) or Van de Ven et al. (2003), relating to five European countries]. Although risk adjustment in Switzerland is based solely on age and sex, it already redistributes a significant amount of money between individuals [regarding the insufficiency of these two criteria, see Van Vliet (2006)]. Its volume amounted to CHF 4.8 billion (bn.) (or some 1 percent of Swiss GDP or 21 percent of payments) as of 2005. This amount of redistribution has given rise to an extensive political discussion. In addition, refining the Swiss volume of risk adjustment scheme has not been debated scientifically.

The present chapter purports to achieve two objectives. First, it investigates the consequence of complementing the risk adjustment formula by a criterion that has been found to be particularly effective [Beck et al. (2006)], viz. "Hospitalization or living in a nursing home during the previous year". Second, it seeks to answer the question of how the volume of risk adjustment could be reduced at minimum cost in terms of increased incentives for risk selection.

Accordingly, the main results are twofold. The introduction of the criterion "Hospitalization or living in a nursing home during the previous year" would inflate the volume of risk adjustment in every canton of Switzerland, in some of them by more than 50 percent. Reducing the volume of risk adjustment (at the national level) from an estimated CHF 5.375 bn. to CHF 4.5 bn. (as of 2005) would reduce cross-subsidization in favor of consumers above 55 by approximately 12 percent while relieving the burden of those below 55 by about 6 percent, using the canton of Zurich as the reference case.

The remainder of this chapter is structured as follows. Section 5.2 is devoted to a description of the data basis, descriptive statistics, and checks of the representativeness of the data. Section 5.3 shows how the volume changes when the additional criterion is introduced. Section 5.4 then takes up the issue of capping the volume of risk adjustment, distinguishing three levels, viz. the level of the nation (level of total volume), the level of insurers, and the level of the insured. The consequences for consumers, the insurers, and Swiss cantons are analyzed. Section 5.5 sets up the optimization problem, while section 5.6 details the opportunity cost of capping risk adjustment, illustrating with an empirical example. Section 5.7 concludes.

5.2 Data basis, descriptive statistics, and representativeness

In order to carry out this research, three large Swiss health insurers provided individual health insurance data. All holders of basic health insurance during the period 2001 to 2005 were considered, totalling 2.78 million (mn.) individuals. Besides socioeconomic variables like age, gender and canton of residence, data on ambulant and stationary health care expenditure (HCE), expenditure for medication, and a variable indicating hospitalization in the previous year were collected. To characterize the type of health insurance, the deductible and a variable indicating choice between conventional and managed care contracts were included as well. With 49.5 percent of women, the sample is well balanced with respect to gender. The market share of the three insurers is stable across the age profile, amounting to 25 percent on average. Across the 26 Swiss cantons, the three insurers are over-represented in eastern and central Switzerland and under-represented in the northern and western parts of the country.

In Swiss health insurance, premiums are community-rated. They are uniform in 16 cantons, the remaining cantons distinguishing high, medium, and low premium regions. In 2005, 32 percent of the population lived in cantons with uniform premiums, while 25 percent lived in a high, 27 percent in a medium, and 16 percent in a low premium region.

With regard to the choice of contract, there is a clear trend toward higher deductibles. Whereas in 2001, policies with the minimum deductible (which amounts to CHF 300, USD 250 at 2007 exchange rates) had a share of over 56 percent, this share had decreased to 50 percent by 2005. The three highest deductibles (CHF 1,500, 2,000 and 2,500) increased in importance from 12 to over 22 percent during the same period.

There is a similar trend in favor of managed-care contracts. Especially consumers aged 31 to 35 use this option, resulting in a share of 22 percent in this age group. However, older age cohorts increasingly prefer managed-care contracts as well. Whereas 18 percent by 2005, only 10 percent of over 80 year old individuals had a managed-care contract in 2001.

In Swiss risk adjustment, only two criteria are considered, age and gender. The age classification comprises 15 classes, starting from 19-25 and continuing in 5 year steps. By law, risk adjustment must not lead to financial reallocation between cantons. The national volume of risk adjustment therefore equals the sum of the cantonal volumes. Computing these volumes using the sample data of the three insurers and their market shares yields a total of CHF 4.13 bn., the national countrywide volume being CHF 4.8 bn. [see KVG (2007)].

In all cantons, the calculated volume falls short of the official one (see Table 5.1). The difference is smallest in the canton of Zurich, Lucerne, and Valais, amounting to less than 10 percent. It is between 10 and 20 percent in 9 other cantons, with the maximum attained in the canton of Basel-Stadt, where the three insurers only hold a very small market share of 9 percent. This marked discrepancy could reflect successful risk selection efforts. Selection efforts have high expected return if targeted at a small population at risk [as shown in Zweifel and Eisen (2003), ch. 5.5]. However, there is no significant (negative) correlation between market share of the three insurers and deviation from the official figure of risk adjustment volume, suggesting that risk selection is not the exposition.

Table 5.2 contains calculated payments per capita for all 30 groups used in risk adjustment, along with their standard errors and official countrywide values. Young men have to pay the highest contributions to risk adjustment (over CHF 2,000 per year), followed by young women with CHF 1,773 per year. Over 90 year old women benefit the most from risk adjustment, to the tune of over CHF 8,600, followed by women of age 86 to 90 with CHF 6,917 and men of age 90+ with CHF 6,731. All age groups over 60 are cross-subsidized by risk adjustment.

A comparison with official values (see the last column of Table 5.2) shows these calculations to be too high for younger and too low for older individuals, especially for women. Insurers receive over CHF 12,000 for a 90+ year old woman and over CHF 9,500 for a man of the same age, while the calculated values are CHF 8,673 and CHF 6,732, respectively. These

Table 5.1: Volume of risk adjustment per canton, 2005

Cant.	mn. CHF*	mn. CHF**	CHF per capita*	CHF per capita**
ZH	735.82	796.47	712.59	771.32
BE	592.03	678.17	767.57	879.25
LU	193.80	211.23	690.70	752.81
UR	17.57	19.97	636.16	723.04
SZ	57.00	70.85	532.71	662.09
OW	13.01	16.41	496.63	626.44
NW	15.13	17.23	484.81	551.93
GL	16.52	19.89	544.36	655.60
ZG	47.57	54.14	565.55	643.69
FR	121.48	146.44	625.25	753.76
SO	122.07	154.65	614.44	778.43
BS	110.21	161.97	718.88	1′056.53
BL	149.81	172.14	696.53	800.34
SH	39.18	47.50	648.89	786.60
AR	20.10	24.85	482.77	596.77
AI	6.48	7.43	569.79	653.53
SG	196.02	238.19	544.53	661.68
GR	95.91	110.26	616.16	708.32
AG	259.61	312.05	573.24	689.02
TH	105.30	126.78	578.77	696.83
TI	226.03	270.02	869.03	1′038.14
VD	430.40	502.39	851.63	994.09
VS	160.39	170.13	684.74	726.33
NE	104.10	130.63	784.08	983.90
GE	257.43	333.04	816.24	1′055.99
JU	39.45	52.32	732.56	971.51

* Sample, ** Official data from KVG (2007). Note: 1 CHF = 0.83 USD (2007)

important deviations are mainly responsible for the underestimation of the total volume of risk adjustment noted above. Table 5.2 also shows that the variance of payments increases with age. While the standard deviation is CHF 494 for young women, it attains to CHF 1,770 for the oldest age class. A likely explanation is that cross-subsidies in favor of the elderly differ more between cantons than those to the detriment the young.

Overall, calculated figures come close enough to official risk adjustment values to justify the use of sample figures in the investigation below. Note that risk adjustment is paid by the insured in the final analysis. Young people who are ’good risks’ pay more than their actuarially fair premium. They subsidize ’bad risks’, mainly elderly people, who pay less than their expected HCE.

Table 5.2: Calculated and official risk adjustment payments per capita according to sex and age, 2005

Men	Average	Std.	Min	Max	Official value
19-25	−2,006.50	505.52	−3006.17	−707.84	−1,963.87
26-30	−1,227.59	833.80	−2165.91	2287.40	−1,889.64
31-35	−900.68	678.91	−1733.38	1202.03	−1,771.42
36-40	−979.03	421.93	−1749.27	247.62	−1,624.49
41-45	−828.69	351.55	−1435.17	−40.31	−1,398.94
46-50	−543.46	465.97	−1615.88	349.08	−1,091.94
51-55	−109.82	378.55	−977.63	714.71	−624.63
56-60	290.34	300.27	−557.57	815.53	13.40
61-65	884.74	418.34	228.53	1648.89	771.06
66-70	1,560.60	598.50	187.69	2464.57	1,638.40
71-75	2,535.19	548.54	982.57	3435.54	2,873.43
76-80	3,208.98	653.35	1884.58	4128.30	3,845.50
81-85	4,127.79	1361.80	1261.52	6983.73	4,986.30
86-90	5,286.51	1208.24	2752.09	7945.75	6,880.09
90+	6,731.78	1513.63	2945.10	8915.78	9,541.96
Women	**Average**	**Std.**	**Min**	**Max**	**Official value**
19-25	−1,772.99	494.20	−2780.08	−974.44	−1,484.37
26-30	−1,024.61	461.54	−2211.50	−311.71	−946.01
31-35	−746.06	559.49	−1694.31	1125.73	−749.83
36-40	−961.00	328.45	−1576.69	−316.11	−924.81
41-45	−965.85	279.05	−1749.34	−535.99	−922.02
46-50	−732.01	309.04	−1295.60	−177.44	−646.82
51-55	−442.87	268.14	−1045.08	106.95	−235.80
56-60	−15.51	321.10	−512.16	841.85	205.36
61-65	443.65	247.14	19.55	764.95	737.31
66-70	981.80	395.53	210.13	1603.77	1,415.39
71-75	1,982.76	446.04	758.34	2662.32	2,385.07
76-80	3,136.84	656.22	1838.10	4406.12	3671.81
81-85	4,641.23	775.55	2788.30	6111.25	5,596.14
86-90	6,917.12	987.66	5115.11	8382.98	8,486.06
90+	8,672.75	1770.15	4464.86	11619.96	12,457.28

5.3 Hospitalization as an additional criterion

Current Swiss risk adjustment uses only the two criteria age and gender. However, hospitalization or living in a nursing home during the previous year is being debated in parliament at the time of writing (2007). Beck (2004) and Beck et al. (2006) estimate that this criterion has considerable predictive power in explaining future HCE. It is very easily collected; moreover, being a dummy variable it does not make computation of risk adjustment payments much more difficult. While the formula currently distinguishes 30 age-sex cells, the number of classes would only increase to 60 [for a discussion on other alternatives and their drawbacks see e.g. Lamers (1999), Van de Ven et al. (2004), Lamers and Van Vliet (2003b), Lamers and Van Vliet (2003a), Ellis and Van de Ven (2000), Beck et al. (2006), and Van de Ven and Schut (2007)].

Taking this additional criterion into account, calculated risk adjustment would increase from CHF 4.13 mn. (as of 2005) to CHF 5.82 mn., i.e. by 40 percent. According to Table 5.3, every canton would exhibit an increase. In absolute value, it correlates positively with the current volume of risk adjustment. It is highest in the most populated cantons, such as Bern, Vaud, and Zurich. In relative terms, the highest increase in volume occurs in the small canton of Obwalden with over 70 percent.

Figure 5.1 illustrates age- and sex-specific effects of the additional risk adjustment criterion in the case of the canton of Zurich. This criterion would cause persons with a hospital stay in 2004 to receive payments in 2005 regardless of age or gender. It is remarkable that until age 50, men receive higher payments (due to a higher rate of hospitalization) than do women. With the additional criterion, women contribute less to risk adjustment and receive higher amounts, with age group 51-55 marking the crossover (the group 66-70 is an exception).

5.4 Limiting the volume of risk adjustment

In view of the rapid increase of the volume of risk adjustment since its introduction in 1996 [KVG (2007)] there is a proposal to limit it to the 2005 value (CHF 4.8 bn.). Indeed, some health insurers derive up to 31 percent of their revenue from risk adjustment payments. This might undermine their incentive to compete for consumers. However, the consequence of any cap is that differences in HCE between high and low risks are not fully compensated

Table 5.3: Risk adjustment without and with the additional criterion, hospitalization or living in nursing home during the previous year, 2005

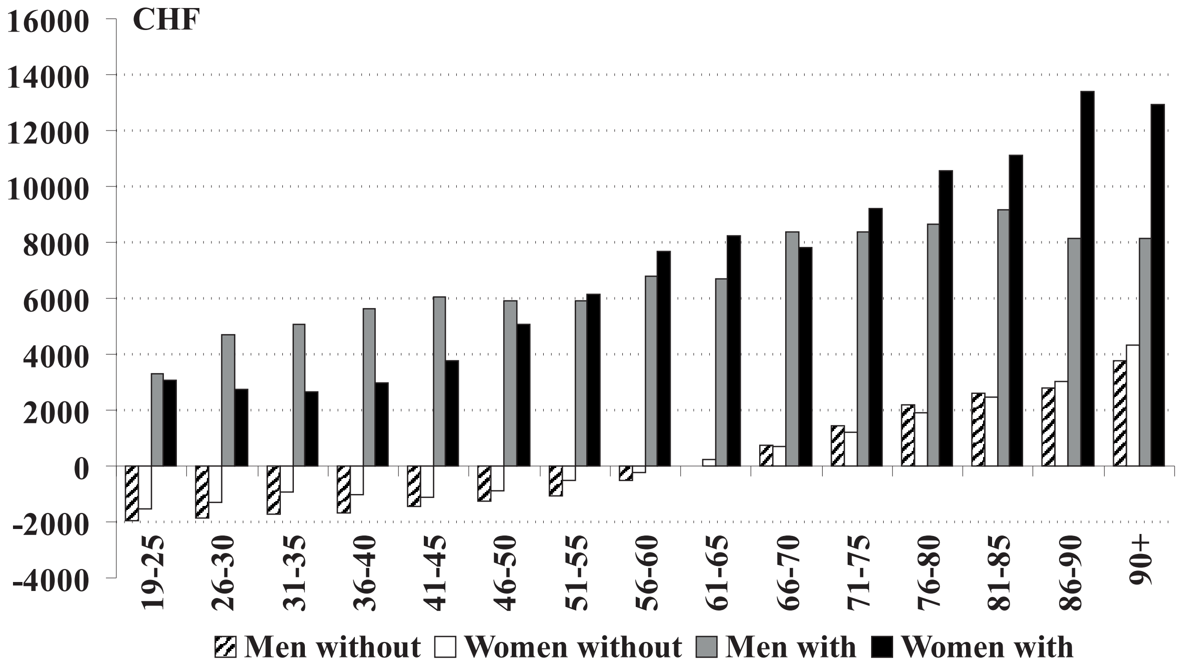
Cant.	mn. CHF*	mn. CHF**	CHF p.c.*	CHF p.c.**	% Increase
ZH	735.82	898.24	712.59	869.89	22.1
BE	592.03	833.26	767.57	1′080.33	40.7
LU	193.80	292.31	690.70	1′041.78	50.8
UR	17.57	26.00	636.16	941.41	48.0
SZ	57.00	84.29	532.71	787.71	47.9
OW	13.37	22.98	496.63	877.33	71.9
NW	15.13	21.35	484.81	683.76	41.0
GL	16.52	24.23	544.36	798.29	46.6
ZG	47.57	71.94	565.55	855.22	51.2
FR	121.48	184.14	625.25	947.79	51.6
SO	122.07	176.52	614.44	888.52	44.6
BS	110.21	171.37	718.88	1′117.84	55.5
BL	149.81	201.91	696.53	938.75	34.8
SH	39.18	59.72	648.89	989.04	52.4
AR	20.10	32.94	482.77	791.23	63.9
AI	6.48	9.58	569.79	842.08	47.8
SG	196.02	298.32	544.53	828.71	52.2
GR	95.91	137.59	616.16	883.88	43.5
AG	259.61	363.64	573.24	802.93	40.1
TH	105.30	173.73	578.77	954.85	65.0
TI	226.03	313.01	869.03	1′203.41	38.5
VD	430.40	593.60	851.63	1′174.55	37.9
VS	160.39	226.85	684.74	968.52	41.4
NE	104.10	166.65	784.08	1′255.26	60.1
GE	257.43	375.23	816.24	1′189.75	45.8
JU	39.45	54.38	732.56	1′009.75	37.8

* without hospitalization/nursing home
** with hospitalization/nursing home

anymore. This results in insurers having to bear additional risk, causing them to engage in risk selection. A natural way to tackle the issue therefore is to cap the volume so that the additional amount of risk borne by insurers is minimized. The limit may be imposed at three levels.

- *At the level of total volume.* The amount to be paid or received by each insurer would be calculated by multiplying the class-specific amount by the number of insured.
- *At the level of insurers.* The individual insurer would know to be paying no more than a certain amount into the risk adjustment scheme (or receive no more than a certain

Figure 5.1: Risk adjustment without / with hospitalization or living in nursing home during the previous year, canton of Zurich, 2005



amount from it). However, payments would need to be adjusted to changes in market structure and enrollment.

- *At the level of the insured.* If payments are limited on an individual basis, premiums would have to be split up in a risk and a redistribution component. Evidencing the reduced redistribution part for low risks and the increased part for high risks, one would achieve maximum transparency.

A cap *at the individual level* would therefore make low risks realize that they contribute to redistribution twice. First, they pay a premium that exceeds their expected cost (solidarity with higher risks). Second, they contribute to the subsidization of premiums of low-income citizens through income taxation (solidarity with lower incomes). High risks with low incomes would have to pay a higher premium than before but would receive more premium subsidies as well. High risks with high incomes would also have to pay more (less cross-subsidization by low risks). Because most politicians prefer hidden redistribution processes to more transparent ones, capping risk adjustment at the individual level is not pursued further.

Capping the volume *at the level of insurers* causes a different set of complications. The two critical issues are the insured population and the number of insurers. Suppose two insurers merge. Some individuals will react to such a merger, either by leaving the merged company or by joining it. A new limit on risk adjustment would need to be calculated in this case. Conversely, insurers could inflate the volume of risk adjustment by splitting up in several different affiliates. Even without any merger activity, the issue remains of how to adjust the caps to inflation, to specific health services inflation, and to technological change in health care.

The last alternative to be examined is *a cap on total volume.* It has the advantage that neither the insured population nor the market structure matters. The problem of adjusting the cap to inflation or technological change remains the same, however. Still, limiting the total volume of risk adjustment seems the most appropriate and plausible way.

A question that naturally arises at this point is who bears the consequences of a cap. Three parties can be identified.

- *The individual insured.* A cap on volume causes premiums to converge towards risk-rated values [with positive effects on moral hazard, e.g. Zweifel and Breuer (2006)].

- *The insurers.* In a system with community-rated premiums, risk adjustment is introduced to eliminate (or reduce) incentives for risk selection. A cap on the volume of risk adjustment benefits insurers with a favorable risk structure while punishing those with an unfavorable one. Not being fully compensated anymore for enrolling unfavorable risks, insurers will try to avoid them. Unfavorable risks then most likely find new cover with an insurer that forms a conglomerate of affiliates, each of which has a uniform premium with one affiliate accepting high risks and the other, collecting the low risks. By placing unfavorable risks in the designated subpool, the risk-selecting conglomerate approaches risk-rated premiums.

- *The cantons.* Capping risk adjustment increases the financial burden of cantons with an unfavorable risk structure. However, the federal government is affected as well because it matches the subsidies paid by cantons. Since premium subsidies are implemented at the cantonal level, generous cantons will attract more high risks and therefore have a larger increase in their subsidies.

5.5 Optimal cap on the volume of risk adjustment

The volume of risk adjustment can theoretically be reduced by an arbitrary amount. One could then analyze the effects on the three parties distinguished in the previous section. Here, the question will be posed the other way round: If the volume of risk adjustment were to be capped, how should this be achieved in an optimal way? Evidently, the opportunity cost in the guise of increased incentive for risk selection should be minimized.

Risk adjustment based on age and gender is calculated in the following way,

$$RA_{a,g} = \bar{L}_{a,g} - \bar{L} \tag{5.1}$$

where a and g are indices for age and gender categories, $RA_{a,g}$ is the payment to ($RA_{a,g} < 0$) or from ($RA_{a,g} > 0$) risk adjustment in group (a, g), $\bar{L}_{a,g}$ is average HCE in group (a, g), and $\bar{L}$ is average HCE in the population as a whole. The volume of risk adjustment can then be calculated as in Equation (5.2),

$$V = \{\sum_{a=1}^{15} \sum_{g=0}^{1} |RA_{a,g}| n_{a,g}\}/2, \tag{5.2}$$

where the division by two avoids double counting. Equation (5.2) is illustrated by a simple example in appendix A.

Risk adjustment is implemented in order to reduce health insurers' incentive to compete for favorable risks. Favorable risks contribute to their margin, which can be used to cover the deficits generated by unfavorable risks. The insurer is exposed to a higher risk of insolvency if these deficits are large. Reserves can be used to ensure solvency, but too many outliers endanger the economic survival of the insurer. There are several methods for analyzing the importance of such outliers, such as value at risk or expected loss at risk [see Hull (2006)]. However, the easiest way to proceed is to analyze the variance.

If one considers age and sex as the only determinants of HCE (which is in accord with current Swiss risk adjustment), then variance in HCE across these groups is given by

$$s_L^2 = \frac{\sum_a \sum_g (\bar{L}_{a,g} - \bar{L})^2 n_{a,g}}{\sum_a \sum_g n_{a,g}} \tag{5.3}$$

where $\bar{L}$ is total average HCE in a specific canton (recall that risk adjustment is calculated in each canton separately), and $\bar{L}_{a,g}$ is average HCE of a specific age and gender cell. As of 2006, s_L^2 is estimated by Beck et al. (2006) at CHF 12 bn.. Risk adjustment thus serves to reduce the variance of HCE falling on insurers (and therefore mitigate the incentive to "skim the cream"). This can be seen by plugging Equation (5.1) into Equation (5.3) and rearranging terms,

$$s_L^2 = \frac{\sum_a \sum_g RA_{a,g}^2 n_{a,g}}{\sum_a \sum_g n_{a,g}} \tag{5.4}$$

and hence,

$$(\sum_a \sum_g n_{a,g}) s_L^2 = \sum_a \sum_g RA_{a,g}^2 n_{a,g}. \tag{5.5}$$

Equation (5.5) shows that with a constant number of individuals in each age and gender cell, risk adjustment payments must be increasing with increasing differences in HCE between groups. This of course serves to increase the volume of risk adjustment as well. If age and gender would be the only determinants of HCE [i.e. insurers have no private information about individuals, contrary to the analysis by Shen and Ellis (2002b), Shen and Ellis (2002a)], then risk adjustment would eliminate all risk induced by community rating. Prior to capping the volume of risk adjustment, the expected variance borne by the health insurer (HI) would then be zero,

$$s_{HI}^2 = 0. \tag{5.6}$$

Capping the volume of risk adjustment causes this variance to become positive. A possible objective would therefore be to minimize its increase, subject to the volume of risk adjustment not exceeding a certain limit. Risk adjustment payments $\widehat{RA}_{a,g}$ are the decision variables in the problem,

$$\min_{\widehat{RA}_{a,g}} \hat{s}_{HI}^2 - s_{HI}^2 \qquad \text{s.t.} \qquad V \leq \bar{V} \tag{5.7}$$

where $\hat{s}^2_{HI}$ means the variance when volume is capped and $s^2_{HI} = 0$ is the variance when the volume is unrestricted.

Of course the optimization must take into account that risk adjustment is zero sum. This however is always achieved since positive and negative payments cancel out.

If volume is defined as in Equation (5.2), optimization is difficult due to absolute values. An alternative approach is therefore taken here. First, the positive half variance $\hat{s}^2_{HI+}$ (with the restriction on volume), and then, the negative half variance $\hat{s}^2_{HI-}$ (with the same restriction on volume) is minimized, ensuring that risk adjustment is zero sum,

$$\min_{\widehat{RA}_{a,g}} [(\hat{s}^2_{HI+}) + (\hat{s}^2_{HI-})] - s^2_{HI}. \tag{5.8}$$

Because $s^2_{HI} = 0$, it is obvious that only the terms in brackets are relevant. The first term can be broken down as shown in Equation (5.8),

$$\min \hat{s}^2_{HI+} = \frac{\sum_a \sum_g (x_{a,g} - \bar{x})^2 n_{a,g}}{\sum_a \sum_g n_{a,g}} \qquad \text{s.t.} \qquad x_{a,g} > \bar{x}. \tag{5.9}$$

The symbols are defined as follows,

$$\begin{aligned} x_{a,g} &= (L_{a,g} - \widehat{RA}_{a,g}) \\ \bar{x} &= \bar{L} \\ x_{a,g} - \bar{x} &= (L_{a,g} - \widehat{RA}_{a,g} - \bar{L}) \\ x_{a,g} - \bar{x} &= (RA_{a,g} - \widehat{RA}_{a,g}). \end{aligned} \tag{5.10}$$

Here, $\widehat{RA}_{a,g} > 0$ if the insurer receives a payment from the risk adjustment and consumers in the (a,g) cell receive a subsidy. Conversely, $\widehat{RA}_{a,g} < 0$ if it pays into the scheme and consumers pay a transfer.

Since the restriction $V \leq \bar{V}$ always holds as an equality in Equation (5.6), the problem can be solved using a Lagrangian,

$$\begin{aligned}\min_{\widehat{RA}_{a,g}} Z &= (\hat{s}^2_+) - \lambda(\sum_a \sum_g \widehat{RA}_{a,g} n_{a,g} - \bar{V}), \qquad \forall \quad \widehat{RA}_{a,g} > 0 \\ &= \frac{\sum_a \sum_g (RA_{a,g} - \widehat{RA}_{a,g})^2 n_{a,g}}{\sum_{a=1}^{15} \sum_{g=0}^{1} n_{a,g}} - \lambda(\sum_a \sum_g \widehat{RA}_{a,g} n_{a,g} - \bar{V})\end{aligned} \tag{5.11}$$

where the subscript HI is dropped for simplicity. Equation (5.11) illustrates how positive payments received from the risk adjustment scheme are optimally reduced. Payments into the scheme ($\forall \quad \widehat{RA}_{a,g} < 0$) are fully analogous.

The first order conditions for this problem are,

$$\begin{aligned}\frac{\partial Z}{\partial \widehat{RA}_{a,g}} &= \frac{-2(RA_{a,g} - \widehat{RA}_{a,g}) n_{a,g}}{\sum_a \sum_g n_{a,g}} - \lambda n_{a,g} = 0 \\ \frac{\partial Z}{\partial \lambda} &= \sum_{a=1}^{15} \sum_{g=0}^{1} \widehat{RA}_{a,g} n_{a,g} - \bar{V} = 0.\end{aligned} \tag{5.12}$$

This is a system of linear Equations in $\widehat{RA}_{a,g}$ and λ that has full rank and can therefore be solved. An example with four risk classes is given in appendix B.

5.6 Consequences of capping risk adjustment

The volume of risk adjustment in Switzerland has increased significantly over the last 10 years. This is illustrated in Table 5.4. The average increase amounts to 6.76 percent per year, attaining CHF 4.864 bn. in 2005. Placing a cap at CHF 5 bn. would be a binding restriction most likely from the year 2006 on. Unless indexed (using e.g. the productivity corrected deflator of HCE), the restriction would become more binding every year, inducing ever stronger incentives for risk selection.

The question as to the optimal value of the cap cannot be addressed in this chapter. It requires knowledge of citizens' willingness to pay for avoiding risk selection efforts by health insurers while keeping community-rated premiums. Experimental evidence concerning willingness to pay for attributes of health services provision has been presented in e.g. Zweifel et al. (2006), Telser et al. (2004), and concerning attributes of health insurance by Becker (2006) and

Table 5.4: Volume of risk adjustment paid between Swiss insured, 1996-2005

Year	Volume (CHF mn.)	Increase
1996	2,700	
1997	2,920	+8.15%
1998	3,195	+9.37%
1999	3,366	+5.32%
2000	3,575	+6.19%
2001	3,826	+7.00%
2002	4,009	+4.77%
2003	4,250	+6.00%
2004	4,568	+7.45%
2005	4,864	+6.44%

Becker et al. (2007). However, willingness to pay for maintaining community-rated premiums has not been measured to the knowledge of the authors. As a second best solution, parliament could decide on the value of the cap, assuming that politicians represent the preferences of the population.

While the political debate has focused on the national level, cantons will likely be affected as well. As evident from Equations (5.10) and (5.12), the opportunity cost of a cap on risk adjustment is linked to the dispersion of HCE in a canton, which varies between cantons. If risk adjustment were to be limited, many citizens with low incomes would have to pay higher premiums. This creates political pressure for increased redistribution through premium subsidies. More generous cantons would be more prone to increasing their subsidies, which are matched by the federal level, where a substantial amount of redistribution between cantons takes place. Therefore, a limit on risk adjustment is likely to induce a certain amount of cross-subsidization between cantons.

A final issue is the impact of an additional criterion (e.g. hospitalization or living in a nursing home during the previous year) on the volume of risk adjustment. First, it is not clear whether the volume increases. If the additional criterion is so effective as to outperform age and sex, then volume would be reduced. This requires the additional criterion to be orthogonal to the existing ones. However, the incidence of hospitalization increases with age and is higher among women than men. Therefore, adding this particular criterion is likely to cause volume to increase. Second, the question arises of whether capping the volume makes this effect more or less important in terms of its opportunity cost (i.e. incentives for risk selection). Generally, three cases can be distinguished.

- The benchmark case is no cap on risk adjustment and the introduction of an additional criterion. This reduces the variance in HCE to be borne by insurers, thus mitigating incentives for risk selection.
- The cap is imposed but not binding initially; additionally it becomes binding with the introduction of the additional criterion. Therefore the opportunity cost of the cap was zero at the beginning. It would become positive but still small if the volume had to be reduced from CHF 5.1 to 5 bn. (say since the effect on insurer behavior is still limited). However, a future reduction from CHF 5.5 to 5 bn. or from CHF 6 to 5 bn. would cause opportunity cost to rise progressively.
- The third alternative is the introduction of an additional criterion when the cap is already binding. On the one hand, this would serve to reduce the net HCE falling on health insurers. On the other hand, the restriction on volume becomes even more binding. The first effect mitigates incentives for risk selection, while the second strengthens them. The net effect remains ambiguous.

Note that the volume of risk adjustment can always be reduced by permitting health insurers to charge premiums that are more in line with true risk. For example, suppose smokers pay an additional premium of CHF 50 per year. This would decrease the difference between HCE and premium revenue by CHF 50 ceteris paribus and hence the variance of payments and hence the risk to be borne by the health insurer. Incentives for risk selection decrease. The advantage of more risk-rated premiums is that the volume of risk adjustment declines endogenously without inducing more efforts at risk selection; its drawback is deviating from community-rating. Conversely, the advantage of capping the volume of risk adjustment is that community-rating can be retained, while its downside is that incentives for risk selection are strengthened.

5.6.1 Empirical illustration

Since the unit of reference for Swiss risk adjustment is the canton, the effects of limiting its volume can be illustrated by using data for the canton of Zurich. Expected HCE was estimated using a two-part model along the lines of Steinmann et al. (2007). However, dummy variables instead of continuous variables were used for age. Other dummies are female sex and canton of residence.

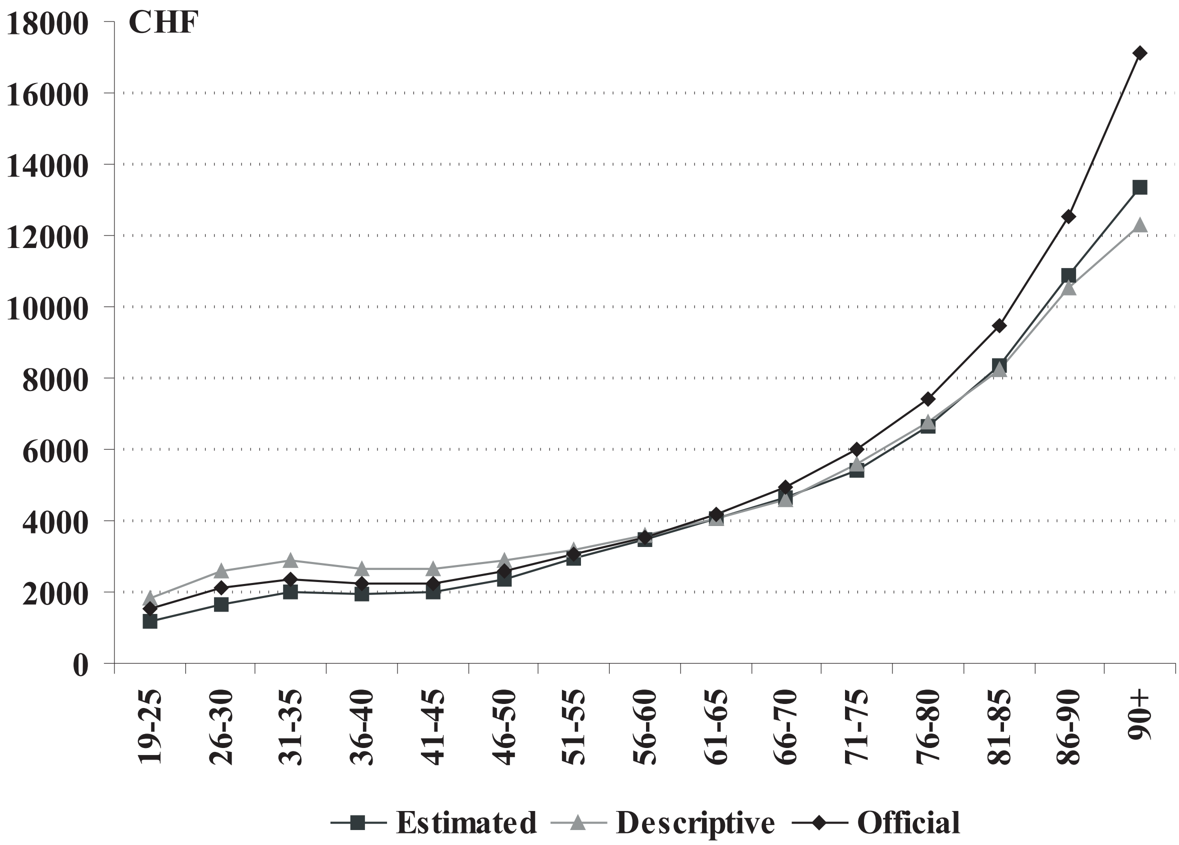
Figure 5.2 shows that the estimated age profile of HCE comes close to that of the three insurers that provided the data. However, the increase of HCE with higher age at the national level is underestimated. Figure 5.3 provides more detail by exhibiting the differences between official average and estimated HCE based on the data of the three insurers. At young ages, HCE is underestimated compared to both the official countrywide average and the three insurers' data. The fit of the Equation is close to perfect between ages 46 and 70. Thereafter, underestimation becomes more important again. Still, estimated HCE seem to fit the Swiss average sufficiently well to derive estimates of the effect a cap on volume may have.

The estimated volume of risk adjustment in its present form is CHF 5.375 bn. between insured. Let this amount be reduced to CHF 4.5 bn.. Table 5.5 illustrates the effects of such a reduction for the canton of Zurich (number of insured: 1,032,600). Initially the youngest age class of women pays a premium that exceeds expected HCE by some CHF 1,600, used to finance the higher HCE of Zurich residents above 50 years of age. Capping the volume of risk adjustment would reduce this excess by CHF 216. On the other hand, the highest age group of females currently receives more than CHF 11,000 from risk adjustment. This cross-subsidy would be reduced by CHF 435. An exception is the class of 51-55 year old women, which changes from receivers to neutral. These amounts are to be compared with the average Zurich premium, which was about CHF 4'000 in 2005.

For insurers, volatility of HCE to be borne increases. Figures 5.4 and 5.5 illustrate estimated risk adjustment payments before and after limiting their volume. For comparison, the current risk adjustment payments are displayed as well. The limitation yields a flattened age profile of risk adjustment payments (as expected). For women the reduction is barely visible in the lower age classes but is maximum at high ages. However, the small number of women in these classes cause total amounts to be negligible.

For men, the picture is slightly different. Below age 55, estimated econometric values consistently overestimate payments into the risk adjustment scheme, and between ages 66 and 75, overestimate payments from the scheme. Again, the very large deviations in the two highest age groups do not matter much because they concern only few individuals.

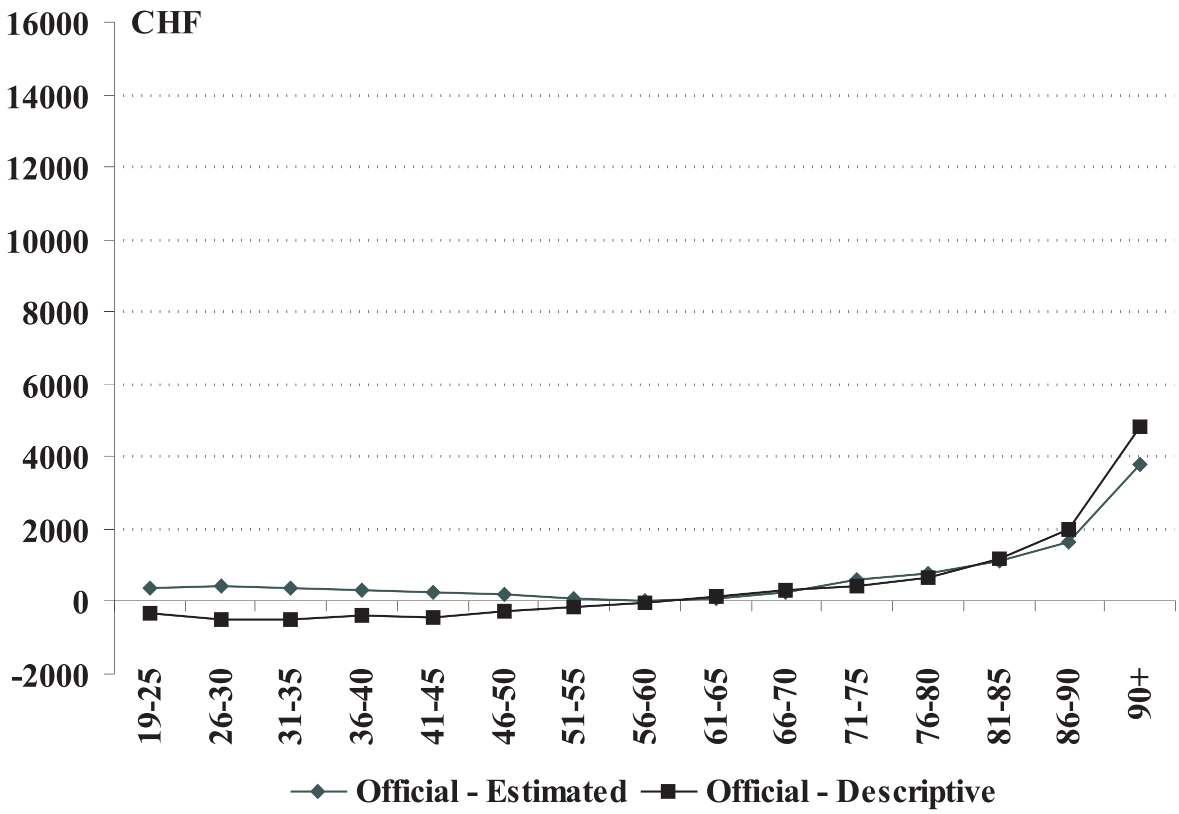
Figure 5.2: HCE of women by age, estimated from 3 insurers and official data, Switzerland, 2005



Naturally, capping the volume of risk adjustment between individuals has an impact on insurers as well. In order to simulate this impact, an additional 22 insurers (covering jointly some 60 percent of the population) were asked to provide aggregate HCE values per age and sex class.

The percentage change between insurers must always be lower than that between insureds since part of the redistribution occurs within the insurance pool. The total volume of risk adjustment between consumers of CHF 5.375 bn. accordingly drops to CHF 1.12 bn. between insurers. Taking the canton of Zurich as the example, this figure is derived as follows. Out of the 1,032,600 insured, 172,671 do not belong to one of the 25 insurers sampled. They hypothetically form one large insurer. This procedure causes the volume of risk adjustment

Figure 5.3: Difference in HCE between estimated from 3 insurers and official data, women, Switzerland, 2005



between insurers to be too low. This underestimation should however not influence the percentage reduction very much (because both before and after values are biased).

The redistribution between insurers operating in Switzerland would drop from an estimated CHF 1.12 bn. to CHF 0.98 bn. or 12.5 percent as of 2005. In an extreme case, insurers may change from paying to receiving, specifically if they mainly enrol 51-55 year old women or 56-60 year old men (see Table 5.5 again).

5.7 Conclusion

This chapter addresses two issues that have been debated in Swiss health policy. On the one hand, the risk adjustment formula may be refined by including an indicator variable,

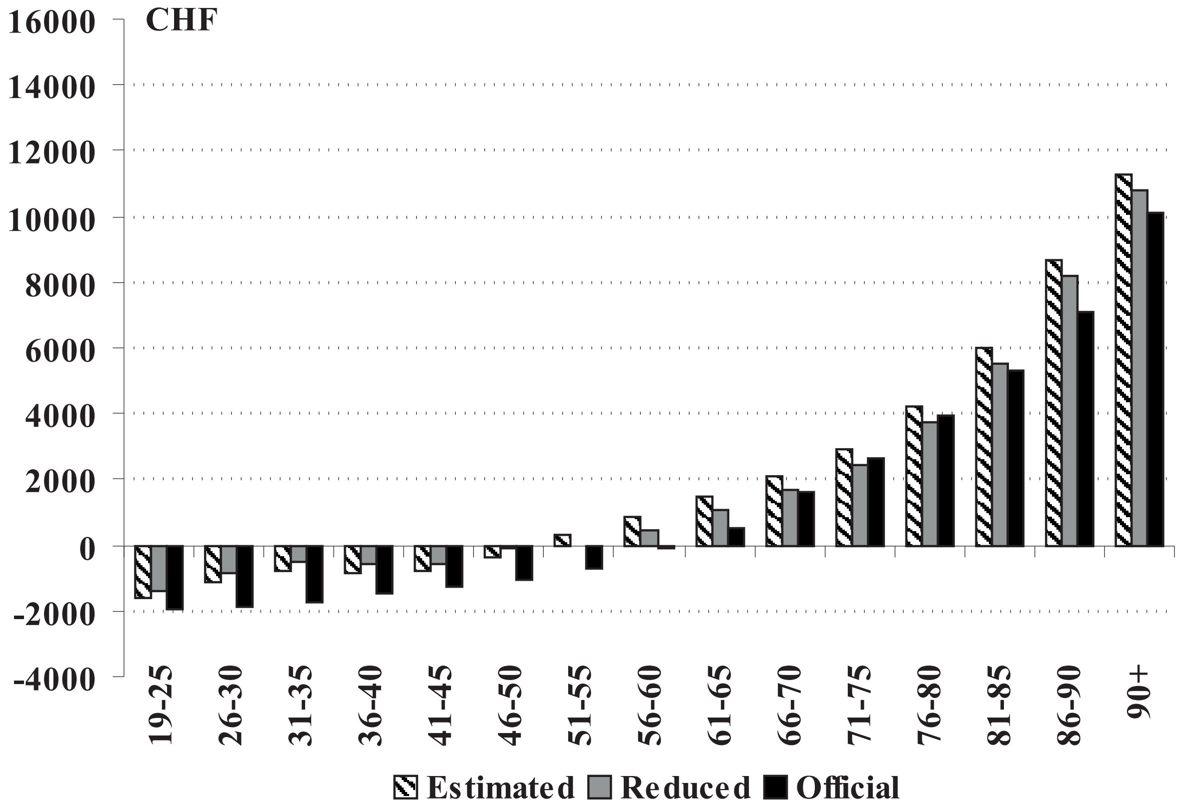
Table 5.5: Capping the volume of risk adjustment; effects at the individual level, CHF, 2005

Women	Amount	Change	New Amount
19-25	−1617.94	+236.01	−1381.93
26-30	−1102.57	+236.01	−866.55
31-35	−771.98	+236.01	−535.96
36-40	−829.99	+236.01	−593.98
41-45	−788.65	+236.01	−552.64
46-50	−349.55	+236.01	−113.54
51-55	283.26	−283.26	0.00
56-60	855.08	−434.50	420.58
61-65	1479.27	−434.50	1044.77
66-70	2125.63	−434.50	1691.13
71-75	2885.94	−434.50	2451.43
76-80	4190.96	−434.50	3756.46
81-85	5983.00	−434.50	5548.50
86-90	8643.28	−434.50	8208.78
90+	11250.27	−434.50	10815.77
Men	Amount	Change	New Amount
19-25	−2347.19	+236.01	−2111.18
26-30	−2289.83	+236.01	−2053.81
31-35	−2191.42	+236.01	−1955.41
36-40	−2072.80	+236.01	−1836.78
41-45	−1800.35	+236.01	−1564.34
46-50	−1478.97	+236.01	−1242.95
51-55	−548.42	+236.01	−312.40
56-60	222.80	−222.80	0.00
61-65	1222.62	−434.50	788.11
66-70	2084.59	−434.50	1650.08
71-75	3300.86	−434.50	2866.36
76-80	3275.79	−434.50	2841.29
81-85	4851.05	−434.50	4416.55
86-90	7161.99	−434.50	6727.48
90+	9419.37	−434.50	8984.86

"Hospitalization or stay in a nursing home during the previous year". On the other hand, as a response to an increase of 80 percent in the volume of risk adjustment since its inception in 1996, there is a desire to cap it. The concern is that a share of up to 31 percent of revenue may undermine insurers' incentive to compete for consumers. However, such a limit has its opportunity cost in the guise of strengthened incentives for risk selection by health insurers.

The present study uses individual data provided by three health insurers. In a first step, they are compared to official nationwide averages to assess their representativeness. The market

Figure 5.4: Estimated, reduced and official risk adjustment payments, women in Zurich, 2005

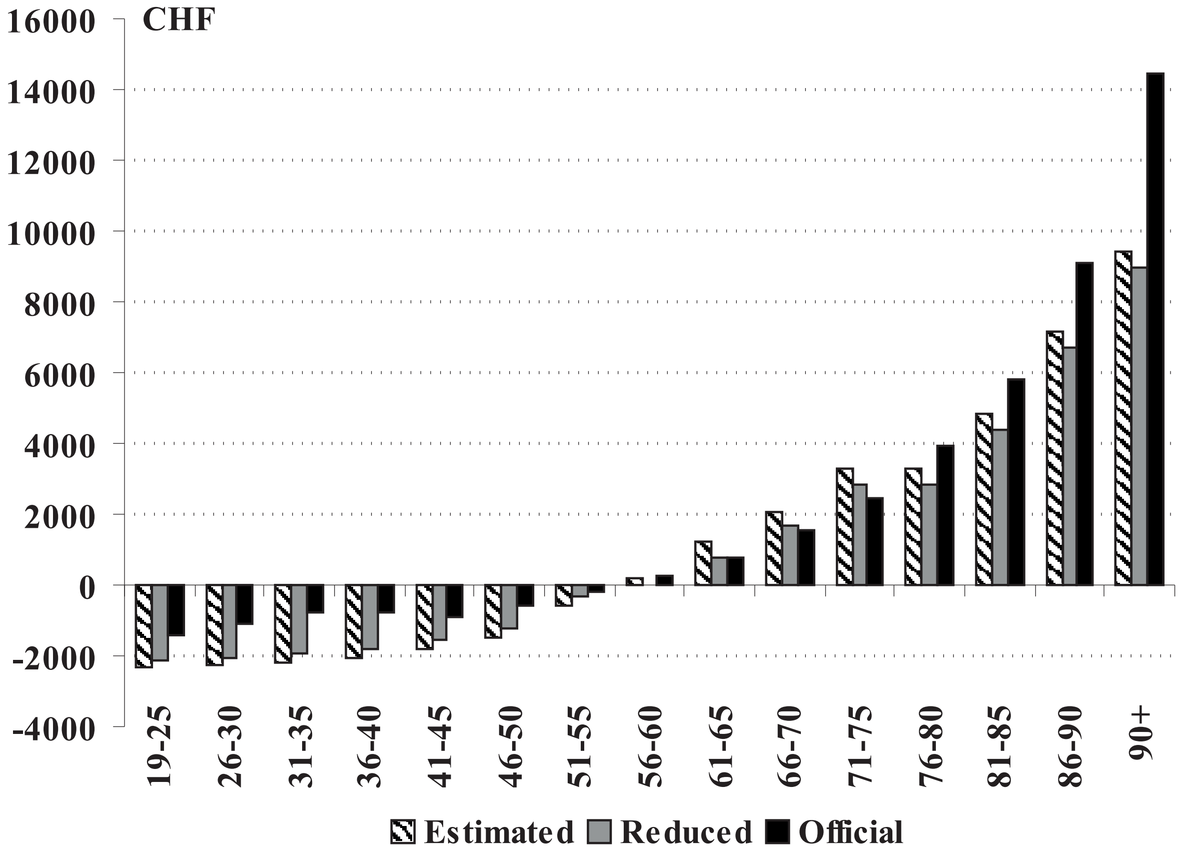


share of the three is above average in the northern and eastern parts and below average in the northwestern and French-speaking parts of Switzerland. Overall, the data seem to accord with official statistics to a sufficient degree to justify more detailed investigation. However, risk adjustment payments in favor of older women are consistently too low resulting in a calculated total volume that is too low.

Inclusion of the additional criterion, “hospitalization or living in a nursing home during the previous year” is found to inflate the volume by 40 percent on average, depending on the canton. Volume would increase most in cantons that currently redistribute a relatively low amount. Age and sex would lose importance as risk adjusters throughout.

A cap on the volume of risk adjustment could be imposed at the nationwide level (capping total volume), the level of health insurers, and the level of individual contribution into and from the scheme by consumers. Limiting total volume turns out to be the most plausible

Figure 5.5: Estimated, reduced and official risk adjustment payments, men in Zurich, 2005



and least complicated solution. There are however implications for consumers (less cross-subsidizing of premiums), the cantons (likely more premium subsidies), and the confederation (more matching grants) and insurers (more efforts at risk selection). The effect on insurers constitutes the optimization problem.

The politically determined objective function is to minimize variance (and therefore the incentive for risk selection), subject to the volume of risk adjustment not exceeding a specified level. The solution is derived by minimizing positive and negative half-variances subject to this restriction. The optimum calls for a uniform reduction of positive and negative risk adjustments.

Table 5.6: Reduction of the volume between insurers

	Before	After	Change
Volume between consumers	5.375 bn.	4.500 bn.	-16.3%
Volume between insurers	1.123 bn.	0.982 bn.	-12.5%

Since an additional criterion is introduced in order to improve the effectiveness of existing risk adjustment the opportunity cost of the cap could considerably decrease. However, this refinement causes volume to increase. Therefore its net effect is ambiguous.

An empirical simulation of the optimal type of cap in one Swiss canton illustrates the argument. It is shown that a reduction to CHF 4.5 bn. (-16 percent) at the national level would reduce the volume of risk adjustment between insurers as well (-12.5 percent). The financial burden drops slightly for those up to age 55; they are juxtaposed by less cross-subsidization in favor of those above 55. The relative change for women is larger than for men throughout. Variance of HCE financed by insurers grows, strengthening their incentives for risk selection.

In all, this research uncovers several uneasy tradeoffs. If community-rating is to be maintained, refining the risk adjustment scheme appears to be desirable in order to mitigate remaining incentives for risk selection. However, this is likely to increase the volume of redistribution even further, which after all forces low risks to contribute more than a fair premium to health insurance, often in spite of their efforts at containing moral hazard. This would call for capping the volume of risk adjustment - which in turn has the opportunity costs in the guise of a strengthened interest in risk selection on the part of health insurers. Moreover, these opportunity costs may well increase when the risk adjustment formula is extended. Future research will have to also take into account the weakening of insurer's incentives for product innovation caused by risk adjustment [as illustrated by Zweifel (2007)].

Appendix

Appendix A

Suppose the population of a Swiss canton consists of one man and one woman of the same age. The man's expected HCE is 1 and the woman's, 1.3. The volume of risk adjustment would therefore be,

$$V = \frac{|RA_M| + |RA_F|}{2} = \frac{0.3}{2} = 0.15, \tag{5.A.1}$$

with RA_M being the risk adjustment payment by the man and RA_F, the payment received by the woman. Replacing RA_M and RA_F by the corresponding expressions from Equation (5.1) yields

$$V = (|L_M - \bar{L}| + |L_F - \bar{L}|) \cdot \frac{1}{2}. \tag{5.A.2}$$

A change in the volume can be obtained by reducing the allowable HCE of women, as shown by the total derivative

$$dV = (| - \frac{\partial \bar{L}}{\partial L_F}| + |1 - \frac{\partial \bar{L}}{\partial L_F}|) dL_F \cdot \frac{1}{2}. \tag{5.A.3}$$

Now impose a reduction of the volume of one third, or 0.05. Therefore, one has

$$dV = -0.05 = (| - 0.5| + |0.5|) dL_F \cdot \frac{1}{2}. \tag{5.A.4}$$

This can be solved to yield

$$\rightarrow dL_F = -0.1. \tag{5.A.5}$$

Thus, the reduction of the volume of risk adjustment by 33 percent is achieved by reducing the allowable HCE of women from 1.3 to 1.2. In other words, only two thirds of the excess HCE of women would continue to be compensated by risk adjustment.

Appendix B

Assume a hypothetical risk adjustment scheme distinguishing four groups with different expected HCE,

$$
\begin{aligned}
n_0 = 3 \quad &; \qquad \bar{L}_0 = 1 \\
n_1 = 4 \quad &; \qquad \bar{L}_1 = 2 \\
n_2 = 2 \quad &; \qquad \bar{L}_2 = 5 \\
n_3 = 1 \quad &; \qquad \bar{L}_3 = 10.
\end{aligned}
$$

This yields the following average HCE, payments to and from risk adjustment for each group, and unrestricted volume,

$$
\begin{aligned}
n &= n_0 + n_1 + n_2 + n_3 \\
\bar{L} &= 3.1 \\
RA_0 &= -2.1 \\
RA_1 &= -1.1 \\
RA_2 &= 1.9 \\
RA_3 &= 6.9 \\
V &= 10.7.
\end{aligned}
$$

Now suppose the objective is to cap the volume at $\bar{V} = 10$ with a minimal increase in the variance that needs to be covered by insurance companies. The first order conditions for negative payments (payments to the risk adjustment scheme) are,

$$
\begin{aligned}
\frac{\partial Z}{\partial \widehat{RA}_0} &= \frac{-2(RA_0 - \widehat{RA}_0)n_0}{n} - \lambda n_0 = 0 \\
\frac{\partial Z}{\partial \widehat{RA}_1} &= \frac{-2(RA_1 - \widehat{RA}_1)n_1}{n} - \lambda n_1 = 0 \\
\frac{\partial Z}{\partial \lambda} &= \widehat{RA}_0 n_0 + \widehat{RA}_1 n_1 - \bar{V} = 0.
\end{aligned}
\tag{5.B.1}
$$

where λ can be solved for from the first FOC,

$$\lambda = \frac{-2(RA_0 - \widehat{RA}_0)}{n}. \tag{5.B.2}$$

Equation (5.B.2) shows the determinants of the opportunity cost caused by a limit on the volume of risk adjustment. First, the greater the population at risk (n), the smaller this cost. Second, the greater the difference between the risk adjustment payment with and without the cap $(RA_0 - \widehat{RA}_0)$, the higher this cost. Alternatively, the system (5.B.1) can be solved to yield,

$$\begin{aligned} (RA_0 - \widehat{RA}_0) &= (RA_1 - \widehat{RA}_1) \\ \widehat{RA}_0 n_0 + \widehat{RA}_1 n_1 &= \bar{V}. \end{aligned} \tag{5.B.3}$$

It is evident that the optimal reduction will always be the same across risk categories. Solving this system of two Equations in the two unknowns yields the following solution for the negative payments (payments to risk adjustment),

$$\begin{aligned} \widehat{RA}_0 &= \frac{\bar{V}^- - RA_1 n_1 + RA_0 n_1}{n_0 + n_1} = \frac{-10 - (-1.1 \cdot 4) + (-2.1 \cdot 4)}{7} = -2 \\ \widehat{RA}_1 &= \frac{\bar{V}^- - RA_0 n_0 + RA_1 n_0}{n_0 + n_1} = \frac{-10 - (-2.1 \cdot 3) + (-1.1 \cdot 3)}{7} = -1. \end{aligned} \tag{5.B.4}$$

Plugging Equation (5.B.4) into the formula for the volume yields $\bar{V}^- = -10$ (the negative value is due to the fact that these are payments into the risk adjustment scheme).

The positive payments (payments from the risk adjustment scheme) can be derived in an analogous way,

$$\begin{aligned} \widehat{RA}_2 &= \frac{\bar{V}^+ - RA_3 n_3 + RA_2 n_3}{n_2 + n_3} = \frac{10 - (6.9 \cdot 1) + (1.9 \cdot 1)}{3} = 1.6\bar{6} \\ \widehat{RA}_3 &= \frac{\bar{V}^+ - RA_2 n_2 + RA_3 n_2}{n_2 + n_3} = \frac{10 - (1.9 \cdot 2) + (6.9 \cdot 2)}{3} = 6.6\bar{6} \end{aligned} \tag{5.B.5}$$

which plugged into the formula for the volume again yields $\bar{V}^{+} = 10$.

Bibliography

Beck, K. (2004). *Risiko Krankenversicherung, Risikomanagement in einem regulierten Krankenversicherungsmarkt - (Risk of Health Insurance, Risk Managment in Regulated Health Insurance).* Bern: Haupt Verlag.

Beck, K., M. Trottmann, U. Kaeser, B. Keller, S. von Rotz, and P. Zweifel (2006). *Nachhaltige Gestaltung des Risikoausgleichs in der Schweizer Krankenversicherung - (Sustainable Design of Swiss Risk Adjustment).* Basel: Ott Verlag.

Becker, K. (2006). *Flexibilisierungsmöglichkeiten in der Krankenversicherung - (Possibilities of more Flexibility in Health Insurance).* Hamburg: Dr. Kovac Verlag.

Becker, K., A. Brändle, and P. Zweifel (2007). *Was wollen die Versicherten? Evidenz aus Deutschland und den Niederlanden - (What do the Insured want? Evidence from Germany and the Netherlands).* Gütersloh: Bertelsmann Stiftung.

Ellis, R. P. and W. P. Van de Ven (2000). Risk adjustment in competitive health plan markets. In A. Culyer and J. Newhouse (Eds.), *Handbook of Health Economics*, pp. 755–845. Amsterdam: North-Holland.

Hull, J. C. (2006). *Options, Futures and Other Derivatives* (6. Edition ed.). New Jersey: Pearson Prentice Hall.

KVG (2007). *www.kvg.org.*

Lamers, L. M. (1999). Risk adjusted capitation based on the diagnostic cost group model: An empirical evaluation with health survey information. *Health Services Research 33*(6), 1727–44.

Lamers, L. M. and R. C. J. A. Van Vliet (2003a). 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. and R. C. J. A. Van Vliet (2003b). The pharmacy-based cost group model: Validating and adjusting the classification of medications for chronic conditions to the Dutch situation. *Health Policy 68*, 113–128.

Shen, J. and R. P. Ellis (2002a). Cost-minimizing risk adjustment. *Journal of Health Economics 21*(3), 515–530.

Shen, J. and R. P. Ellis (2002b). How profitable is risk-selection? A comparison of four risk adjustment models. *Health Economics 11*, 165–174.

Steinmann, L., H. Telser, and P. Zweifel (2007). Aging and future healthcare expenditure: A consistent approach. *Forum for Health Economics and Policy 10*(2), Article 1.

Telser, H., S. Vaterlaus, P. Zweifel, and P. Eugster (2004). *Was leistet unser Gesundheitswesen? - (What does our Health System achieve?)*. Zürich: Rüegger Verlag.

Van de Ven, W., K. Beck, F. Buchner, D. Chernikovsky, L. Gardiol, A. Holly, L. Lamers, E. Schokkaert, A. Shmueli, S. Spycher, C. Can 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., R. V. 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. and F. Schut (2007). Risk equalization in an individual health insurance market: The only escape from the tradeoff between affordability, efficiency and selection - the Netherlands as a case study. *Working Paper*.

Van Vliet, R. C. J. A. (2006). Free choice of health plan combined with risk-adjusted capitation payments: Are switchers and new enrolees good risks? *Health Economics 15*, 763–774.

Zweifel, P. (2007). The theory of social health insurance. *Foundations and Trends in Microeconomics 3*(3), 183–273.

Zweifel, P. and M. Breuer (2006). The case for risk-based premiums in public health insurance. *Health Economics, Policy and Law 1*(2), 171–188.

Zweifel, P. and R. Eisen (2003). *Versicherungsökonomik - (Insurance Economics)*. Heidelberg: Springer Verlag, 2. Auflage.

Zweifel, P., H. Telser, and S. Vaterlaus (2006). Consumer resistance against regulation: The case of health care. *Journal of Regulatory Economics 29*(3), 319–332.

Can Guaranteed Renewability Survive in the Presence of Death?

Patrick Eugster*

Will be submitted to the "Journal of Risk and Uncertainty".

*The author would like to thank his thesis supervisor Peter Zweifel (University of Zurich) for valuable discussions, comments and inputs.

Chapter 6

Can guaranteed renewability survive in the presence of death?

6.1 Introduction

This chapter starts by noting that there are basically two solutions for dealing with risk selection (i.e. high-risk individuals remaining uninsured) in long-term health insurance. One is to impose community rating by law; the other is to make contracts incentive-compatible. The latter, market-based solution has been investigated by Pauly, Kuhnreuther and Hirth (1995), PKH henceforth, and Cochrane (1995) in their seminal contributions. They derive long-term insurance contracts that exhibit a feature of time consistency. In PKH this sequence of incentive-compatible short-term contracts results in guaranteed renewability (GR). As envisioned by Cochrane (1995, p. 447) these contracts must be renegotiation-proof, satisfying a participation constraint in that all parties are always willing to sign the next contract under all future circumstances. GR contracts achieve time-consistency by frontloading the sequence. As pointed out by Frick (1998), however, frontloading may be excessive for individuals who have limited capital endowment. Yet, the empirical evidence presented in Herring and Pauly (2006) suggests that this is not a problem in reality.

The present chapter adds to the debate by complementing the set of transition probabilities as follows. Whereas Pauly et al. (1995) consider only the likelihood of a favorable risk turning to an unfavorable one during the planning period, the possibility of returning to favorable risk status is accounted for as well. It will be shown that this causes frontloading to be less than in the PKH model, with the amount of reduction increasing in the probability of returning to favorable risk status. Premiums depend on the difference between expected cost pertaining to a weighted average of high and low risks and that pertaining to low risks. Moreover, when the Markov process is complemented by death as a terminal state, the amount of frontloading is reduced even more.

These statements are supported by a hypothetical comparison between the premium derived in this chapter and the illustration presented by PKH. Moreover, it is shown that once death and the cost of dying are included in the model, the critical value of time preference that is still compatible with GR decreases causing demand for GR contracts to increase. In conclusion, GR can be said not only to survive in the presence of death but to mitigate the problems of adverse selection, cream skimming, and risk rating described by Patel and Pauly (2002). In private health insurance, demand for (unrestricted) GR insurance contracts is enhanced once the underlying Markov model is specified more completely.

The remainder of this chapter is structured as follows. The next section reviews the literature on guaranteed renewable and long-term insurance. Section 6.3 presents the model and derives the predictions, which are then used to compare the critical value of time preference derived by Frick (1998) with the value that follows from the completed model (section 6.4). Section 6.5 illustrates a more general approach. Section 6.6 concludes.

6.2 Related Literature

There are two ways to counter risk selection and its associated problems in long-term health insurance, with risk selection defined as high-risk individuals remaining uninsured. One is to introduce mandatory community-rating as is the case in employer-contracted health insurance in the United States and individually-contracted health insurance in Switzerland. The other is to have risk-averse consumers pay a surcharge for GR. This second solution is prevalent in the U.S. individual health insurance market. Note that this requires health insurers to have the right to risk rate premiums.[1] Risk-averse consumers buy health insurance voluntarily, with the premium development ensuring time consistency. Patel and Pauly (2002, p. 283) describe the policy choice in the following way: *The alternative to guaranteed renewability,*

[1] Interestingly, Herring and Pauly (2001) show that in fact premiums vary more strongly with risk in community-rated areas than in risk-rated areas of the United States.

for people concerned about adverse selection, risk rating, or cream skimming, is, as Paul Ginsburg notes, 'setting strong (regulatory) rules for this market.' Those rules usually entail some kind of community rating or limits on risk rating. Such rules themselves cause adverse selection and cream skimming, so they tend to beget still more restrictions on the kinds of policies that can be offered.

However, GR is unlikely to work perfectly either. In particular, individuals typically pay less than under perfect risk rating, presumably because insurers fear the bad publicity 'excessive premiums' may trigger. Therefore, some adverse side effects of imperfect risk rating remain, e.g. a lock-in of high risks who want to make do with a GR contract [Patel and Pauly (2002)]. Still, GR is a market-based mechanism that at least partially overcomes the problem of uncertainty about long-term health status.

The theory of long-term health insurance was initially developed by PKH and Cochrane (1995). Long-term insurance contracts are defined as a sequence of short-term contracts that are incentive compatible, in accordance with the participation constraint and renegotiation-proof, as stated in the prior section. Through applying game theory and backward induction PKH derive an incentive-compatible contract sequence with declining premiums. GR thus may be a way to ensure long-term health insurance for an entire population, provided however everyone is actually capable and willing to prepay the frontloading in GR insurance. Cochrane (1995), however, uses a multi-period utility function in discrete time. His solution is an additional account from which so-called bidirectional severance payments (equal to the increased present value of premiums or expected future cost) avoid ex-post defection by one party (this can also be an individual who gets healthier unexpectedly).

The starting point of this chapter is the optimal contract in PKH, who deduces a time-consistent premium schedule from expected cost development over time. The more the planning period is extended into the far future, the larger becomes the initial prepayment to ensure GR. At least in the United States, the longest observable time horizon is defined by eligibility for Medicare (age 65). In countries without a scheme similar to Medicare, the planning period is limited by the expected time of death. In these countries, transition between health states must include death as well in an attempt to derive a time-consistent optimal health insurance contract.[2]

However, consider a U.S. worker who enters the private insurance market at age 20 and remains in it until the age of 65. Even in this case prepayment to ensure GR can be a heavy financial burden and even outright unaffordable for young individuals in the presence of credit market imperfections. Frick (1998) shows that imperfect capital markets in combination with differing subjective rates of time preference result in some consumers failing to purchase a GR contract. Indeed, a critical value of time preference exists with consumers who are below dropping out of GR.

Assuming that the number of high-loss periods is fixed and the same for everyone, Pauly et al. (1998) derive a level premium schedule for group health insurance (e.g. employer-contracted). In fact, a double pooling mechanism can be created (at the group level and at the insurer level, through combining several groups with the same insurer). Obviously, such a double pooling serves to reduce frontloading for GR [Pauly et al. (1998)]. The level premium schedule should of course be sustainable even in the case of credit market imperfections [not noted by Pauly et al. (1998)]. However, all it takes to obtain it is to assume a positive interest rate paid on the frontloading surcharge and a uniform premium rebate at the end for all those remaining in the GR pool. This double pooling mechanism may already work in reality, since Herring and Pauly (2001) analyzing individual records found that premiums are not fully risk rated in the U.S. market for individually contracted health insurance even in states where it is permitted.

Two other recent empirical contributions deal with GR contracts. Brown III and Conelly (2005a) evaluate the Australian government's initiative to foster long-term private health insurance using a GR model with probabilities published in Herring and Pauly (2006). They find that Australia's lifetime cover is subject to adverse selection the way it is set up currently. However, GR may constitute a voluntary alternative to lifetime cover because it avoids penalizing loading in later years of life, since GR premiums contain an upfront surcharge instead. In a second contribution, Brown III and Conelly (2005b) extend the PKH model to 35 periods, allowing for age dependent loss probabilities. They hypothesize that the existence of large

[2] Death should be endogenously modelled to depict reality even in the case where such a system exists [see a short discussion on endogenizing death in Steinmann et al. (2007)]. Whenever a system like Medicare does not exist, however, modelling death is even more important.

cross-subsidies from healthy, younger individuals to less healthy, older ones is one of the key factors preventing the young from voluntarily buying health insurance. Moreover, they argue that by substituting to prepayment, GR contracts may overcome market failure resulting in consumers lacking cover for a future deterioration of their risk status.

Finally Herring and Pauly (2006) estimate the amount of frontloading in existing GR health insurance in the United States. They construct a risk-rated GR premium profile and compare it to the observed development of premiums during the life of the contract. They find an amazing degree of similarity between the predicted and the actual time paths in U.S. individually contracted private health insurance. While acknowledging the importance of their assumptions in obtaining this result, they fail to relax them where possible.

This chapter seeks to remedy this, mainly by introducing death as the 'absorbing state' (in Markovian language). This important modification of course raises the question of whether the GR concept survives in the face of death. The finding is that the presence of death actually enhances the feasibility of GR because it causes the upfront GR surcharge to be lower.

6.3 The Model

The main element for calculating a lifetime health insurance premium that is compatible with GR is a set of transition probabilities (see Figure 6.1). As in PKH, there is a probability p_{12} of an insured turning from a low risk (with expected future cost $p_l L$, where p_l is the common loss probability and L the size of the loss) to a high risk (characterized by $p_h \geq p_l$ and/or $H \geq L$, respectively). Accordingly, p_{11} is the probability of remaining in low-risk status.

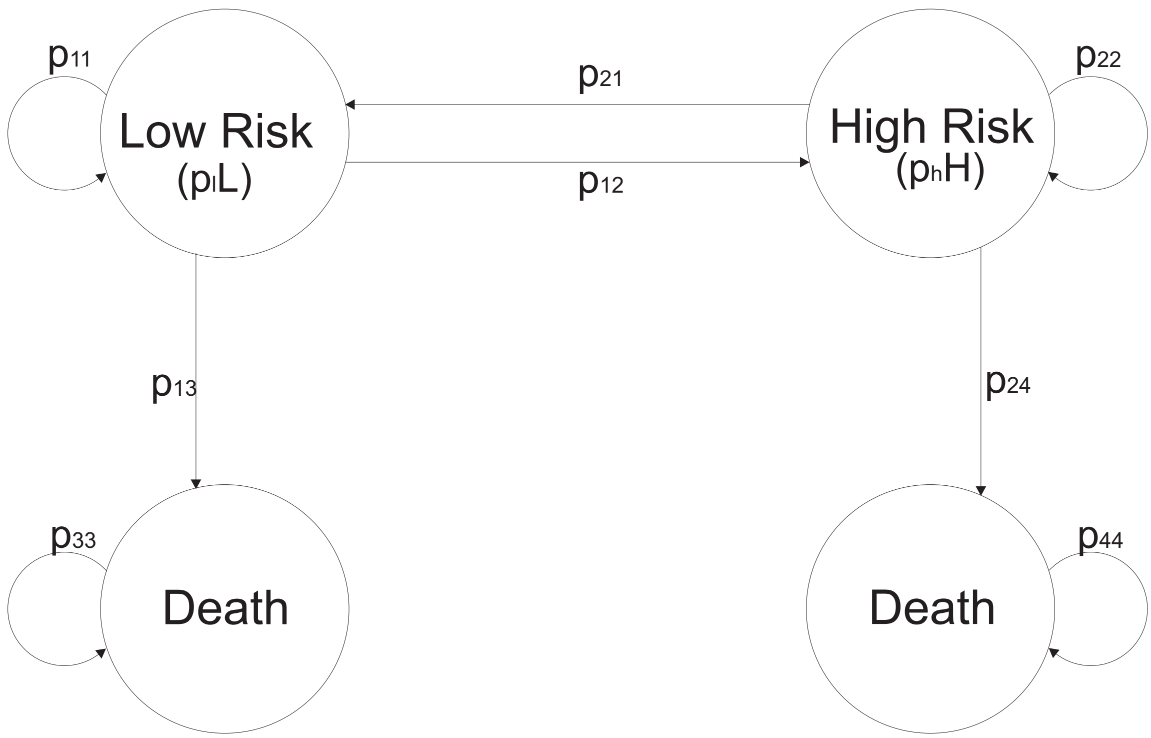
Unlike PKH, the model contains important additional probabilities. First, high risks are assumed to return to low-risk status with probability $p_{21} > 0$. Accordingly, they also have a probability (p_{22}) of remaining high risks. However, they may die ($p_{24} > 0$), which is also true of low risks, with probability $p_{13} > 0$.[3] The transition probabilities are collected in Matrix A,

[3] In a recent empirical contribution, Herring and Pauly (2006) present evidence suggesting that the probability to die for high and low risks differs significantly. To keep matters simple the graphical presentation could only model death once with an average probability to die for the whole population. This, however, would not depict reality and therefore both states of death are modelled. While both states may differ in terms of probability of death and its respective cost engendered, they are both absorbing ($p_{33} = p_{44} = 1$).

$$A = \begin{bmatrix} p_{11} & p_{12} & p_{13} & 0 \\ p_{21} & p_{22} & 0 & p_{24} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{6.1}$$

Clearly, PKH deal with a special case of Equation (6.1), viz. $p_{21} = p_{13} = p_{24} = 0$ and $p_{22} = 1$, while the loss probabilities in the low risk state are $p_{11} = 1 - p_l$ and $p_{12} = p_l$, respectively. These restrictions are relaxed step by step below.

Figure 6.1: Underlying Markov process



6.3.1 Positive probability of returning to favorable risk status

Neglecting the two absorbing states of death but assuming $p_{21} > 0$, the transition matrix A of Equation (6.1) reduces to Matrix B,

$$B = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}. \tag{6.2}$$

In order to compare the results with PKH calculations, their values $p_{11} = 0.9$ and $p_{12} = 0.1$ are adopted here. Moreover, $p_{21} = 0.25$ (fixing p_{22} at 0.75) is assumed.[4]

Table 6.1 displays the development of the share of high and low risks in the PKH model and in its modification. The loss probability is derived by weighting the state-dependent loss probabilities [(p_l) and (p_h)] with the respective share. Intuitively a positive probability of returning to favorable risk status should lower long-run premiums and hence the amount of frontloading. This is illustrated in Table 6.1 using the same parameter values as PKH (details are provided in appendix A). Appendix A also shows that premiums depend on the difference between expected cost pertaining to a weighted average of favorable and unfavorable risks and that pertaining to favorable risks.

During the first two periods, the modification $p_{21} = 0.25 > 0$ does not have an effect. However, from then on the share of high risks approaches 0.1248 rather than 0.3439. The loss probability begins to change in period three already, converging to 0.12496 rather than 0.16878. In combination the lifetime premium amounts to $P_{Mod.} = 59.376$ rather than $P_{PKH} = 68.908$, a reduction of 13.8 percent. Note that a two-period model would fail to indicate this change, being subject to the restriction that the premium of the last period must be less or equal to the actuarially fair premium for a low-risk individual. Otherwise, low risks would not take out insurance. However, because the last-period premium and the probability of transition from low to high risk status are given while $p_{21} > 0$ is not relevant yet, the second to last period premium is determined as well and cannot differ from the value calculated by PKH.

[4] In a binary distribution with probabilities π, $(1 - \pi)$, the mean waiting period for transition from state 1 to state 2 is $D = 1/\pi$, see e.g. Bhattacharyya and Johnson (1977). Pauly et al. (1998) assume $D = 4$ in their GR group insurance model, therefore $\pi = p_{21} = 0.25$.

Table 6.1: PKH model and modification

Period	PKH → $p_{21} = 0$			Modified → $p_{21} = 0.25$		
	High Risks	Low Risks	Loss Pr.	High Risks	Low Risks	Loss Pr.
1	0.0000	1.0000	0.10000	0.0000	1.0000	0.10000
2	0.1000	0.9000	0.12000	0.1000	0.9000	0.12000
3	0.1900	0.8100	0.13800	0.1200	0.8800	0.12400
4	0.2710	0.7290	0.16230	0.1240	0.8760	0.12480
5	0.3439	0.6561	0.16878	0.1248	0.8752	0.12496
			$P_{PKH} = 68.908^*$			$P_{Mod.} = 59.376^*$

* No discounting for simplicity

Therefore, in order to see the effect of $p_{21} > 0$, the Markov process must go on for more than two periods.

Equation (6.2) can be manipulated to establish a relationship between PKH and the GR group insurance model by Pauly et al. (1998). The steady-state distribution vector v of an ergodic Markov process satisfies,

$$v' \times B = v'. \tag{6.3}$$

The following system of Equations therefore solves for the steady state distribution vector v,

$$L + H = 1 \tag{6.4}$$

$$\begin{bmatrix} L & H \end{bmatrix} \times \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} L & H \end{bmatrix} \tag{6.5}$$

Solving for L and H yields,

$$L = \frac{p_{21}}{1 + p_{21} - p_{11}} = 0.714 \tag{6.6}$$

$$H = \frac{p_{12}}{1 + p_{12} - p_{22}} = 0.286. \tag{6.7}$$

In the long run 71.4 percent of the insured belong to the favorable and 28.6 percent to the unfavorable risk category, respectively.[5]

6.3.2 Introducing Death I: Time of Death

At this point, death is reintroduced as an absorbing state. However, the time of death proves to be very important. In analogy to PKH, the argument always assumes that the individual starts in the low-risk state at time t, while the analysis stops at time $t+1$.

Figure 6.2 (a) illustrates the PKH timing. Once the premium is paid, cost may accrue. Then risk status may change, before period $t+1$ starts. Accordingly, premiums must satisfy the following restriction in a two-period model,

$$P_t + P_{t+1} = p_l L + p_{11} p_l L + p_{12} p_h H. \tag{6.8}$$

A one-period contract would require the premium for low risks to be no higher than their expected cost. Therefore P_{t+1} can be set equal to $p_l L$. This in turn implies that the premium in period t is given by,

$$P_{t,a} = p_l L + p_{12}(p_h H - p_l L), \tag{6.9}$$

where $P_{t,a}$ is the premium in period t with the timing illustrated in Figure 6.2 (a) and $p_{11} = 1 - p_{12}$ since $p_{13} = 0$. $P_{t,a}$ is the premium derived in PKH.

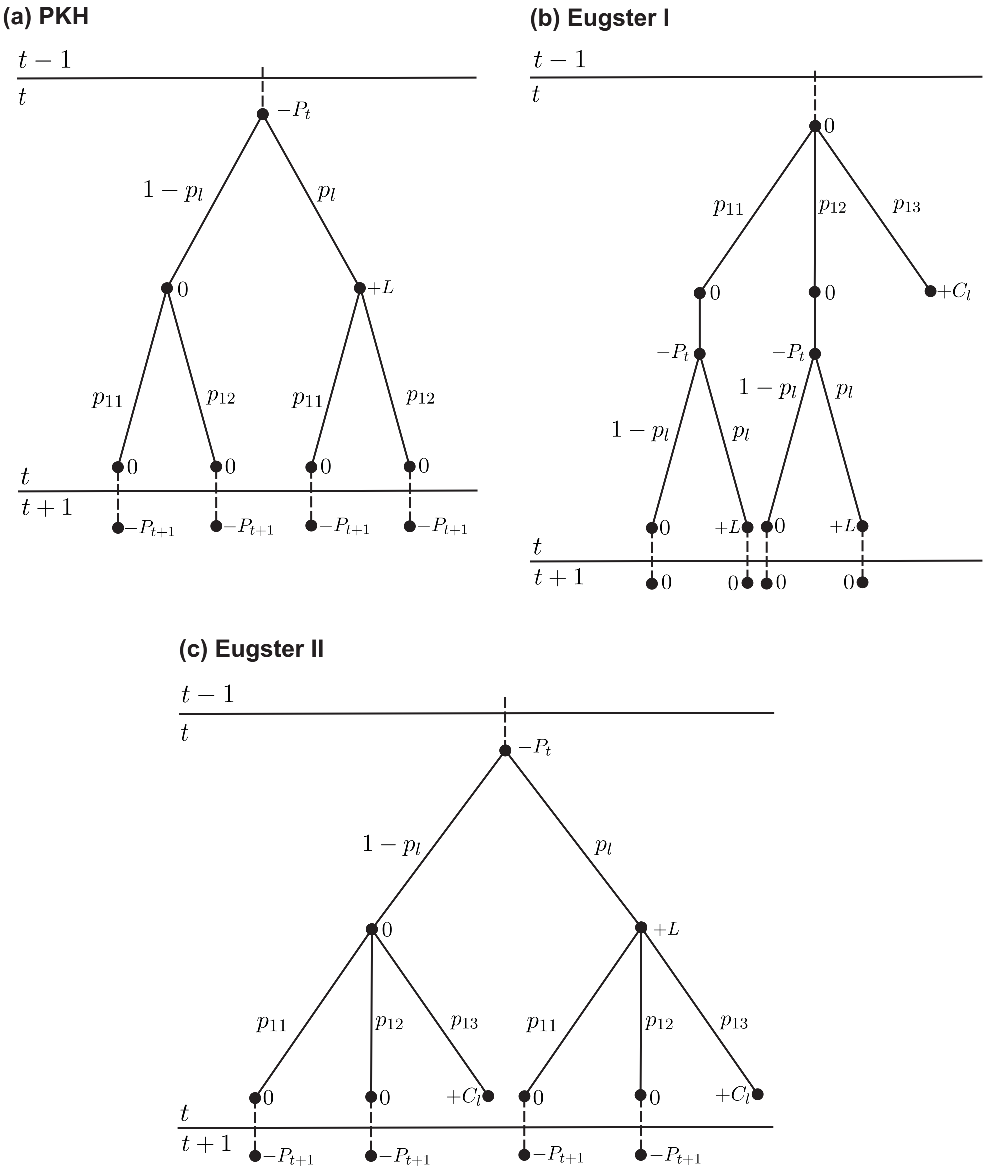
Now death is introduced. Let the change in risk status take place before the premium is paid [see Figure 6.2 (b)]. Because death and change in risk status occur at the beginning of the period, only low risks are offered GR insurance in period t and only survivors pay the premium in period $t+1$. Premiums must now satisfy,

$$P_t + (1 - p_{13})P_{t+1} = p_l L + p_{11} p_l L + p_{12} p_h H + p_{13} C_l, \tag{6.10}$$

where C_l is the extra cost of dying as a low risk. Again substituting $P_{t+1} = p_l L$ and rearranging terms, one has,

[5]Pauly et al. (1998) as well as Nickel (2005) derive the same steady-state distribution. Level contributions are then calculated from this steady state.

Figure 6.2: The crucial importance of time of death



$$P_{t,b} \quad = \quad p_{11}p_lL + p_{12}p_hH + p_{13}(C_l + p_lL) \tag{6.11}$$

$$P_{t,b} \quad = \quad p_lL + p_{12}(p_hH - p_lL) + p_{13}C_l, \tag{6.12}$$

since $p_{11}+p_{13} = 1-p_{12}$. Three possible scenarios emerge from introducing death as a terminal state:

- A positive probability of death ($p_{13} > 0$) *reduces* the probability of remaining in the low-risk status but leaves the probability of transition from low- to high-risk status *unchanged* ($p_{11} \downarrow$, $p_{12} \rightarrow$).
- A positive probability of death ($p_{13} > 0$) *reduces* the probability of transition from low- to high-risk status but leaves the probability of remaining in the low-risk status *unchanged* ($p_{11} \rightarrow$, $p_{12} \downarrow$).
- A positive probability of death ($p_{13} > 0$) *reduces* both the probability of transition from low- to high-risk status and the probability of remaining in the low-risk state ($p_{11} \downarrow$, $p_{12} \downarrow$).

Figure 6.2 (c) introduces a third possible modification. Here, transition occurs at the end of the period after cost accrued. In analogy to Figure 6.2 (b) only survivors pay the premium in period $t+1$. However, everyone is offered a contract in period t since the change in risk status occurs at the end of the period after the premium is paid. This timing implies the following set of restrictions on the premium,

$$P_t + (1 - p_{13})P_{t+1} = p_l L + p_{13}C_l + p_{11}p_l L + p_{12}p_h H + p_{11}p_{13}C_l + p_{12}p_{24}C_h, \tag{6.13}$$

where C_h is the extra cost of dying as a high risk. Now P_{t+1} must also cover the extra cost of dying as a low risk. The restriction $P_{t+1} = p_l L + p_{13}C_l$ is imposed on Equation (6.13). This implies that the premium in period t is reduced to,

$$\begin{aligned} P_{t,c} &= p_{11}p_l L + p_{12}p_h H + p_{13}(p_l L + p_{13}C_l) + p_{11}p_{13}C_l + p_{12}p_{24}C_h & (6.14) \\ &= p_l L + p_{12}(p_h H - p_l L) + p_{13}C_l + p_{12}(p_{24}C_h - p_{13}C_l). & (6.15) \end{aligned}$$

In full analogy to Figure 6.2 (b), three scenarios can be distinguished. Note that both premiums including death are lower than or equal to PKH as long as zero excess cost of dying is assumed ($C_l = C_h = 0$). The first two terms on the RHS of Equation (6.15) serve to cover expected health loss, with the first term referring to a low risk during period t and the second, to the possible increase in expected cost weighted with the probability of becoming a high risk in period $t+1$. Up to this point, Equation (6.15) is in full analogy of the PKH model.

With ($C_l > 0$ and $C_h > 0$), however, there are two additional terms that are due to death as a terminal state. One of them is for the expected extra cost of dying in t as a low risk, the other, for the possible increase or decrease in the extra cost of dying due to a transition from low- to high-risk status at the end of period t.

6.3.3 Introducing Death II: Extra Cost of Dying

During the past few years, there has been a growing literature revolving around the 'red herring' that claims the influence of age on health care expenditure to be dwarfed by that of closeness to death [e.g. see Lubitz and Riley (1993), Zweifel et al. (1999), Werblow et al. (2007), or Steinmann et al. (2007)]. This subsection seeks to elucidate the impact of the extra cost of dying ($C_h, C_l > 0$) on the GR premium, again using Figure 6.2.

As to Figure 6.2 (a), the PKH case, the cost of dying is not considered. Figure 6.2 (b) and Equation (6.12) again constitute a more complex case and all three scenarios [defined below Equation (6.12)] can be distinguished. Scenario one clearly increases the premium in period t while scenarios two and three depend on the relationship between the cost of dying as a low risk and the difference in expected cost between low and high risks.

Finally, turn to Figure 6.2 (c). In the first scenario, $P_{t,c}$ clearly increases compared to $P_{t,a}$. Turning to scenario two and three one can distinguish no less than nine subcases. The upshot is that there are many parameter constellations that cause the GR premium with death to be lower in the first period than calculated by PKH ($P_{t,a}$). This in turn implies that critical time preference that is still compatible with GR insurance is lowered (i.e. generating a higher demand for GR insurance contracts as shown in section 6.5.2).

6.4 Credit Rationing Revisited

As stated in the literature review, Frick (1998) argues that consumers with limited income may be subject to credit rationing in imperfect capital markets, rendering them incapable of financing the GR surcharge upfront. In this section, his argument is revisited, using the different chronologies developed in Figure 6.2 and the corresponding premium structures. Following Frick (1998) this multi-period model considers two periods only and assumes logarithmic utility. The optimization problem then reads (with y denoting constant per period income to finance premiums),

$$\begin{aligned} \max_{P_1, P_{2H}} \quad & U(y - P_1) + p_{12}\beta U(y - P_{2H}) + p_{11}\beta U(y - P_{2L}) \\ \text{s.t.} \quad & P_1 + p_{12}P_{2H} + p_{11}P_{2L} = p_l L + p_{11}p_l L + p_{12}p_h H \\ & P_{2L} = p_l L \end{aligned} \tag{6.16}$$

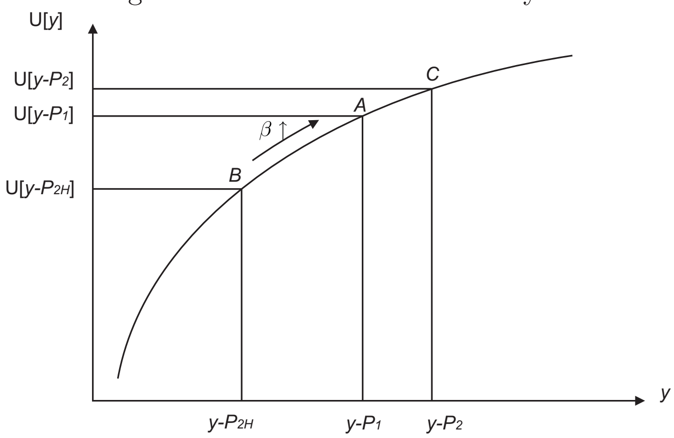
Here, P_1 is the first-period premium paid by everyone as a low risk, while the second-period premium differs between risks. The second-period premium paid by those who become high risks P_{2H} differs from that paid by low risks P_{2L}, with $\beta \leq 1$ being the subjective discount rate. The Lagrangian and its first-order conditions yield,

$$U'[y - P_1] = \beta U'[y - P_{2H}] \tag{6.17}$$

$$p_l L + p_{12}p_h L = P_1 + p_{12}P_{2H}. \tag{6.18}$$

Note that consumers can attain these conditions only given the assumption that a new insurer may enter the market in any period. An optimization over just two periods is therefore sufficient. It is revealed that the higher the time preference β of individuals, the higher is the premium P_{2H} that individuals are willing to accept. Figure 6.3 illustrates.

Figure 6.3: Consumer's risk utility function



Let the risk utility function of a consumer who has paid P_1 and hence has first-period disposable income $(y - P_1)$ pass through point A. Given that the consumer remains in the contract and pays premium P_{2H} from his second period income, let the risk utility function pass through point B, with higher marginal utility. Now with $\beta = 1$ (no time preference), points A and B would have to coincide to make marginal utilities equal in both periods. However, if marginal utility $U'[y - P_{2H}]$ is multiplied by $\beta < 1$ (indicating positive time preference), point A must lie to the right of point B, indicating that a second-period premium $P_{2H} > P_1$ is acceptable in optimization. As one can glean from Equation (6.18), P_{2H} only enters the budget constraint of individuals if $p_{12} > 0$, i.e. a positive probability of becoming an unfavorable risk. Therefore, Equations (6.17) and (6.18) together imply that the insurers have some scope in skimming off willingness to pay by consumers who have both marked time-preference and a positive probability of becoming a high risk.

The formula derived by Frick (1998) can be generalized to include additional probabilities and manipulated to derive long-term premium paths, permitting to calculate it for any contract length. However, the major contribution by Frick (1998) is to prove the existence of a critical value of time preference $\bar{\beta}$ separating individuals who buy GR insurance from those who do not. Based on logarithmic utility and the actuarially fair premium per period, one obtains these threshold values of $\bar{\beta}$ for all three chronologies in Figure 6.2,

$$\bar{\beta}^a = \frac{y - p_h H}{y - p_l L} \qquad \text{(as in PKH);} \tag{6.19a}$$

$$\bar{\beta}^b = \frac{y - p_h H}{y - p_l L - p_{13} C_l}; \tag{6.19b}$$

$$\bar{\beta}^c = \frac{y - p_h H - p_{24} C_h}{y - p_l L - p_{13} C_l}. \tag{6.19c}$$

Compared to PKH [Equation (6.19a)], chronology (b) results in a higher rather than lower critical value of time preference. A lower value of $\bar{\beta}$ is only possible in chronology (c). However, the outcome depends on the relative magnitude of $p_{24}C_h$ and $p_{13}C_l$. As pointed out in section 6.3, Herring and Pauly (2006) estimate the probability of death to be significantly larger for high risks. The cost of death with regard to different risks has never been studied econometrically and would be an interesting topic for future research.

6.5 A More General Approach

Alternatively the optimization problem can retain the second-period, community-rated premium. A consumer considering a future change in risk status now optimizes,

$$\begin{aligned} \max_{P_1,P_2} \quad & U(y-P_1)+(1-p_{13})\beta U(y-P_2) \\ \text{s.t.} \quad & P_1+(1-p_{13})P_2 = p_l L + p_{11}p_l L + p_{12}p_h H \\ & P_2 \leq p_l L \end{aligned} \tag{6.20}$$

where exemplary the constraint in form of Equation (6.8) is used. The corresponding first order conditions are,

$$U'[y-P_1] = \beta U'[y-P_2] + \frac{\mu}{(1-p_{13})} \tag{6.21}$$

$$P_1+(1-p_{13})P_2 = p_l L + p_{11}p_l L + p_{12}p_h H \tag{6.22}$$

Two cases can be distinguished,

- $P_2 = p_l L \;\rightarrow\; \mu \geq 0$. This would imply $U'[y-P_1] \geq \beta U'[y-P_2]$.
- $P_2 < p_l L \;\rightarrow\; \mu = 0$. This would imply $U'[y-P_1] = \beta U'[y-P_2]$.

In the first case, the constrained second-period premium prohibits the financial burden on consumers from increasing later (see Figure 6.3 again, with $\beta = 1$, the optimal value for consumers could even be a point like C, to the right of point A). The second case is similar to the result derived by Frick (1998), where consumers want to equate first- and second-period marginal utilities. This means that premium P_{t+1} exceeds P_t and would be the same again only in the special case of $\beta = 1$ (A and C coincide in Figure 6.3).

6.5.1 Positive Rate of Interest and Increasing Health Care Cost

Up to this point, the main question to be answered has been whether GR contracts are likely to be sustainable in the presence of death. While the answer tends to be positive with regard to the timing presented in Figure 6.2 (c), there are additional issues confronting policy makers. On the one hand there is the unrelenting increase in the cost of health care which poses a challenge to any health insurer seeking to write GR contracts. On the other hand increases in the cost of health care relax the second-period premium constraint. In addition, financing the upfront GR surcharge would be more affordable if consumers' payments were credited with a positive rate of interest. With $g > 0$ denoting the rate of cost increase and $i > 0$ the rate of interest paid on the GR frontloading, the budget constraints [Equations (6.8), (6.10), and (6.13)] become,

$$(1+i)P_t + P_{t+1} = (1+i)p_l L + p_{11} p_l L(1+g) + p_{12} p_h H(1+g) \tag{6.23a}$$

$$\begin{aligned}(1+i)P_t + (1-p_{13})P_{t+1} = {} & (1+i)p_l L + p_{11} p_l L(1+g) + p_{12} p_h H(1+g) + \\ & p_{13} C_l (1+g)\end{aligned} \tag{6.23b}$$

$$\begin{aligned}(1+i)P_t + (1-p_{13})P_{t+1} = {} & (1+i)p_l L + (1+i)p_{13} C_l + p_{11} p_l L(1+g) + p_{12} p_h H(1+g) + \\ & p_{11} p_{13} C_l (1+g) + p_{12} p_{24} C_h (1+g)\end{aligned} \tag{6.23c}$$

The time of death and its impact on the critical rate of time preference $\bar{\beta}$ is analyzed in the next section. Generally, a positive probability of death continues to lower $\bar{\beta}$, thus causing demand for GR insurance contracts to increase.

6.5.2 Determining the Critical Value of Time Preference

The optimization problem composed of Equation (6.20) and restrictions (6.23a)-(6.23c) are now examined and a numerical example will be used to illustrate the influence of time of death, continuing and comparing the three chronologies introduced in section 6.3.2.

To keep the results comparable to Pauly et al. (1995) and Frick (1998), Table 6.2 first contains the same parameter values, with the probability of falling ill (p_l) set equal to the probability of transition from low- to high-risk status (p_{12}) and the probability of remaining in high-risk status set equal to the probability of positive health care cost in the high-risk state ($p_h = p_{22}$). The calculation proceeds as follows. Suppose $\beta = 1$. Unconstrained individuals equate marginal utilities across periods, as stated below Equation (6.22). In order to check whether being unconstrained is possible (with $\beta = 1$), the maximum premium that a low-risk status consumer is willing to pay in $t + 1$ is substituted into the restrictions ensuring that premi-

Table 6.2: Parameter values assumed

Chronology (a), no death (PKH)						Chronologies (b) and (c), with death					
p_{11}	=	0.9	p_h	=	0.25	p_{11}	=	0.9	L	=	20
p_{12}	=	0.1	L	=	20	p_{12}	=	0.07	H	=	20
p_{13}	=	0	H	=	20	p_{13}	=	0.03	Y	=	100
p_{21}	=	0.75	Y	=	100	p_{21}	=	0.65	i	=	0.1
p_{22}	=	0.25	i	=	0.1	p_{22}	=	0.25	g	=	0.1
p_{24}	=	0	g	=	0.1	p_{24}	=	0.1	C_l	=	5
p_l	=	0.1				p_l	=	0.1	C_h	=	2
						p_h	=	0.25			

ums cover expected cost. Inserting the values given in Table 6.2 into Equation (6.23a), one obtains[6],

$$(1+i)P_{t,a} + P_{t+1} = (1+i)p_l L + p_{11} p_l L(1+g) + p_{12} p_h H(1+g)$$
$$P_{t,a} = 2.3.$$

Every consumer is constrained in chronology (a), because $P_{t,a} > P_{t+1} = p_l L(1+g) = 2.2$. The lowest possible value of time preference compatible with unconstrained optimization can be derived by dividing first-period by second-period marginal utility, with both premiums set to their actuarially fair per-period values (the fair per-period premium in $t+1$ is the low risks' expected cost and any remaining costs are paid in period t),

$$\frac{U'[Y - P_{t,fair}]}{U'[Y - P_{t+1,fair}]} = \bar{\beta}. \tag{6.24}$$

The RHS of Table 6.2 displays the parameter values used for the two chronologies (b) and (c) that deal with death. Assuming one quarter high-risk individuals (approximately the steady-state distribution of section 6.3.1), average death probability in this population is $p_{Death} = 0.047$.[7] For chronology (b) they are plugged into Equation (6.23b),

[6] An alternative way of arriving at the same result is to set premiums in both periods equal and solve for a uniform premium across periods using the budget constraint. If the result exceeds the maximum possible premium in the later period, then even the most patient individuals ($\beta = 1$) are restricted in this optimization.

[7] Life Expectancy at birth for Swiss men is 77.2 years with average probability of death at age 77 being 0.049 [see Federal Statistical Office (BFS) (2007)]. Since GR may be used as a lifetime contract this end of life setting is assumed to model the last two periods.

$$(1+i)P_{t,b} + (1-p_{13})P_{t+1} = (1+i)p_l L + p_{11}p_l L(1+g) + p_{12}p_h H(1+g) + p_{13}C_l(1+g)$$

$$P_{t,b} = 2.236.$$

This value again exceeds the maximum premium a low-risk consumer is willing to pay in the second period, viz. 2.2 as before. Turning to chronology (c) one has,

$$(1+i)P_{t,c} + (1-p_{13})P_{t+1} = (1+i)p_l L + (1+i)p_{13}C_l + p_{11}p_l L(1+g) + p_{12}p_h H(1+g) + p_{11}p_{13}C_l(1+g) + p_{12}p_{24}C_h(1+g)$$

$$P_{t,c} = 2.363.$$

This falls short of the maximum acceptable value of $P_{t+1} = 2.365$. Table 6.3 summarizes the effect of each parameter on both premiums [in chronology (c)] and states whether an increase in the respective parameter helps to achieve $P_{t+1} \geq P_t$.

Table 6.3: Derivatives of both premiums and their effects

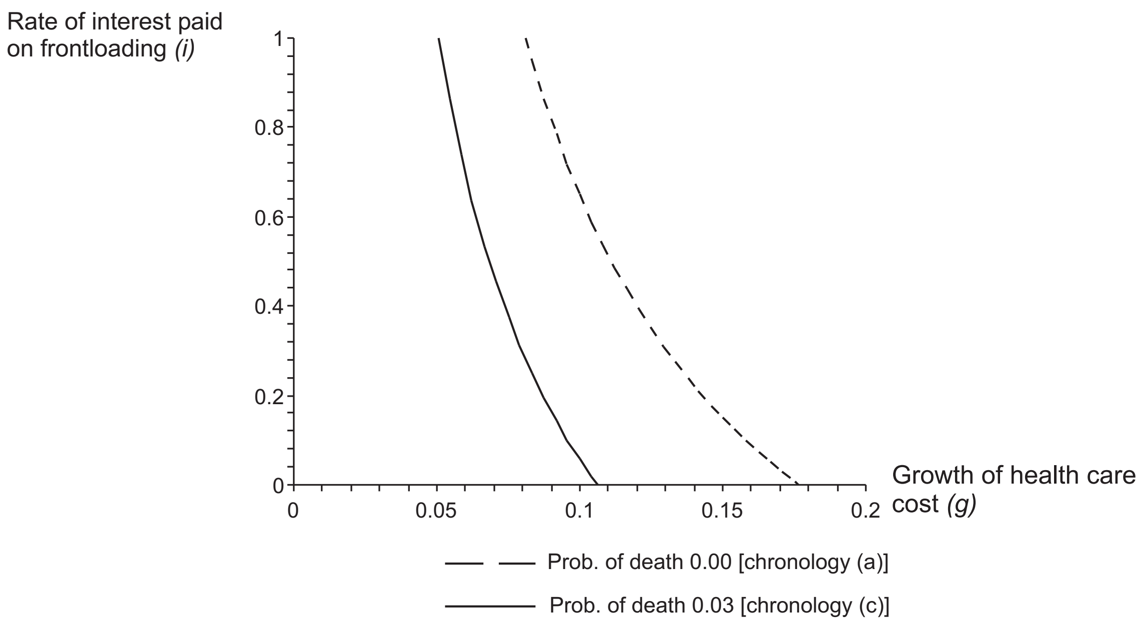
Parameter	$\frac{\partial P_t}{\partial \text{Parameter}}$	$\frac{\partial P_{t+1}}{\partial \text{Parameter}}$	Effect
i	$-Rp_{12}((p_h H - p_l L) + (p_{24}C_h - p_{13}C_l))/(1+i)$	0	✓
g	$p_{12}((p_h H - p_l L) + (p_{24}C_h - p_{13}C_l))/(1+i)$	$p_l L + p_{13}C_l$	(✓)
C_l	$p_{13}(1 - p_{12}R)$	$p_{13}(1+g)$	✓
C_h	$p_{12}p_{24}R$	0	–
$p_l L$	$1 - p_{12}R$	$1+g$	✓
$p_h H$	$p_{12}R$	0	–
p_{24}	$p_{12}C_h R$	0	–

Note: (✓) = (contingent on parameter values) increasing this parameter relaxes the constraint; – = increasing this parameter tightens the constraint; $R = (1+g)/(1+i)$.

On the one hand this is true whenever P_t is increased less than P_{t+1}. On the other hand this occurs when P_t is decreased more than P_{t+1}. Increasing the rate of interest paid on prepayments, expected health care costs, or the cost of dying as a low risk all relax the second-period premium constraint. Increasing the rate of growth in health care costs just does so if $p_{12}(p_h H + p_{24}C_h) < (1+i+p_{12})(p_l L + p_{13}C_l)$. However, every parameter associated with the high-risk state aggravates the second-period premium constraint. The outcome of changing a transition probability in the low-risk state crucially depends on the remaining parameters and the effect on its composite probabilities (see appendix B).

More generally, there is a critical combination of cost growth (g) and interest rate (i) that can achieve uniform premiums over two periods.[8] The locus of critical values in Figure 6.4 is derived by assuming that the second- and the first-period premium are the same. Figure 6.4 first displays the PKH case (zero probability of death). Any combination of (g, i) lying to the origin from this locus cannot achieve uniform premiums. This is only possible whenever growth in the cost of health care exceeds 17.8 percent. If the insurer earns positive interest on the GR frontloading, this critical value drops close to 10 percent. However, the steep slope of the locus makes clear that the easying element is cost growth (g) since it directly affects the restriction on the second-period premium. The second locus of Figure 6.4 corresponds to chronology (c) (positive probability of death). With the parameter values given, (g) may be below 11 percent p.a. and uniform premiums are possible. Increasing the rate of interest in this case lowers the respective growth rate to almost 5 percent.

Figure 6.4: Combination of i and g that are compatible with GR



[8] Any individual with $\beta = 1$ would prefer uniform premiums, while those with $\beta < 1$ would prefer $P_t < P_{t+1}$. If uniform premiums are not viable because $P_{t+1} < P_t$ always holds every individual will always be constrained.

Another locus of interest is the net gain in expected utility from a GR contract as a function of time preference β. The expected utility gained from GR insurance less the present value of the associated penalty through the restricted second-period premium with parameters again taken from the RHS of Table 6.2 is displayed in Figure 6.5 [deriving Figure 6.5 is illustrated in appendix C using the chronology from Figure 6.2 (c)]. As could be expected, consumers with a high level of time preference ($\beta < 0.526$) would rather have no insurance than a GR contract since the utility reduction outweighs the benefit from insurance. However, there is a group of (rather patient) consumers who benefit from a GR contract. On the other hand, imposing a GR mandate probably would result in an overall efficiency loss.

Figure 6.5: Net expected utility of GR insurance as a function of time preference

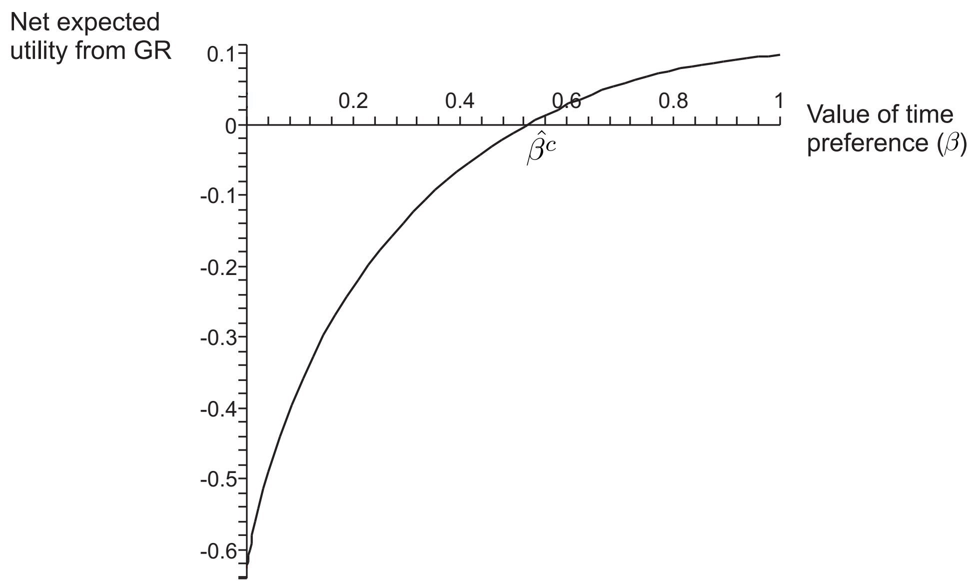
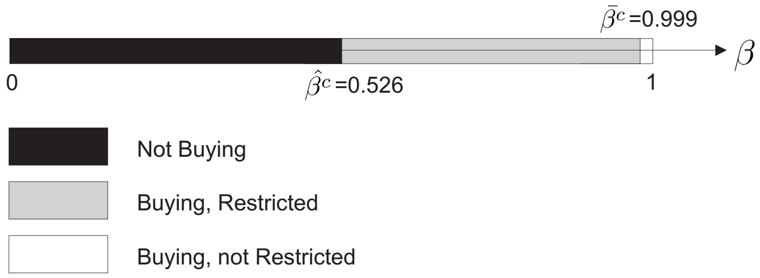


Figure 6.5 derives the lower critical value ($\hat{\beta}^c$) for chronology (c) and is drawn for consumers with Lagrange multiplier $\mu \geq 0$ in the optimization problem [Equation (6.20)]. Therefore, the constraint on their second-period premium is binding. If this constraint is relaxed, one obtains a second, upper critical value ($\bar{\beta}^c$) for chronology (c). This value may be compared with the value that would result from Equation (6.19c). The critical value in the model by Frick (1998) is lower because the community-rating of P_{t+1} is retained here.

Figure 6.6 displays the critical values of the 'more general approach' that are based on the same calibrations as Figures 6.4 and 6.5. One can distinguish the following three types.

- Individuals with strong preference for current consumption $\beta < 0.526$ will not buy GR insurance because the GR surcharge contained in their first-period premium exceeds their expected utility gain from a GR contract.
- Individuals with a moderate to low preference for current consumption $0.526 \leq \beta < 0.999$ do buy GR insurance but would prefer shifting more of the premium burden to the later period.
- Very patient individuals $0.999 \leq \beta \leq 1$ also opt for the GR contract. They would not even have wanted to shift more of the premium burden to the later period.

Figure 6.6: Upper and lower critical values of time preference



This third set would be empty if death had not been considered (as in PKH and depicted in Figure 6.4 where the point $i = g = 0.1$ is to the left of the locus without death) because this serves to shift critical $\bar{\beta}$ values up towards 1. Therefore, GR not only survives but may actually thrive in the presence of death.

6.6 Conclusion

This chapter takes the contribution by Pauly, Kuhnreuther and Hirth (1995) – PKH – on guaranteed renewability (GR) as its starting point. However, PKH implicitly assume that there is only one possible and permanent transition, from low-risk status to high-risk status. This neglects two important possibilities. First, some insureds do revert back to low-risk status, as has been documented e.g. by Beck and Zweifel (1998) for an observation period of five years and taken into account by Pauly et al. (1998) in their calibration of GR contracts in U.S. group health insurance. Second, there is always the transition to death that has been neglected up to now in any theoretical contribution. Therefore, the question arises of whether an individual GR insurance contract 'survives' these modifications or becomes a concept void of practical relevance.

One way to answer this question is to determine the value of time preference that is just compatible with GR. Consumers who discount the future heavily do not want to pay the GR surcharge in the guise of frontloading because they are strongly interested in current consumption, while patient consumers tend to opt for it. If the critical value distinguishing these two types is increased by the two modifications considered, then the set of GR purchasers shrinks, possibly to an empty set. If it is lowered, they serve to enhance GR in health insurance.

However, this test turns out not to be sufficient in all cases. Notably, including the probability of transition from high- to low-risk status does not affect this critical value of time preference while still serving to reduce lifetime premiums by 13.8 percent, thus making GR more attractive in the example given in section 6.3.1. When death is accounted for as a terminal state, the critical value of time preference is not affected in all cases either. The outcome strongly depends on the time of death and its respective cost. This is illustrated for the model by Frick (1998) and the alternative approach (section 6.5.2). Thus, GR is found to survive well in the presence of death. This is also true when the cost of health care increases and interest is being paid on the frontloading provided by consumers. However, the rate of interest turns out to be far less important. Of course, these statements are conditional on the parameter values entering simulations. Analyzing the effect of the remaining parameter values using the more general approach reveals that results are most sensible to changes in low risks' cost of health care, their probability and their cost of death. Future research should therefore also address the empirical question of the ratio between different risks' cost of dying. In addition, the consequence of using alternative utility functions should be studied.

Finally, consumers may face even harsher credit rationing than assumed here when having to come up with the GR surcharge early in their life cycle. Nevertheless, this research permits to conclude that GR constitutes a possible solution to the problem of insuring the risk of deteriorating health status without any particular mandate by public policy.

Appendix A

The optimal premium schedule is derived along the lines of the PKH model. However, the possibility of transition from high- to low-risk status is taken into account [cost of dying is dealt with later; therefore, $(C_l = C_h = p_{13} = p_{24} = 0)$]. The example calculated is limited to five periods since Equations increase in complexity without adding insights (indeed, a three-period formulation suffices, as shown below and stated by PKH who also demonstrate the solution to be a subgame perfect equilibrium, with insurers always wanting to offer the pertinent contract).

Expected cost is given by:

$$EC_t = Prob_{low} EC_{low} + Prob_{high} EC_{high}; \tag{6.A.1}$$

where EC is expected cost, $Prob_{low}$ is the probability of being in the low-risk status t periods from now (defined analogously for the high-risk status).

The corresponding state probabilities can be calculated by raising the transition matrix A to its various powers and assuming an initial distribution in the form of row vector (a). The second-period probability values are therefore,

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}}_{=a} \cdot A \cdot A = a \cdot A^2$$

$$= \left[\underbrace{(p_{11}^2 + p_{12}p_{21})}_{Prob_{low}} \quad \underbrace{(p_{12}p_{22} + p_{11}p_{12})}_{Prob_{high}} \quad \underbrace{(p_{11}p_{13} + p_{13})}_{Prob_{deathlow}} \quad \underbrace{(p_{12}p_{24})}_{Prob_{deathhigh}} \right].$$

where probabilities are defined in Equation (6.1) of the main text.

Expected cost is calculated for low and high risks respectively as indicated by Equation (6.A.1), e.g. $EC_{low} = p_l \cdot L$ with p_l denoting the probability of having positive cost and L, average health care cost pertaining to a low risk. Expected cost for high risk is defined in the same way.

Turning to the five-period formulation, one obtains

$$
\begin{aligned}
EC_1 =& p_l L && (6.A.2)\\
EC_2 =& p_{11} p_l L + p_{12} p_h H && (6.A.3)\\
EC_3 =& (p_{11}^2 + p_{12} p_{21}) p_l L + (p_{12} p_{22} + p_{11} p_{12}) p_h H && (6.A.4)\\
EC_4 =& (p_{11}^3 + 2 p_{11} p_{12} p_{21} + p_{12} p_{21} p_{22}) p_l L && (6.A.5)\\
& + (p_{11}^2 p_{12} + p_{12}^2 p_{21} + p_{12} p_{22}^2 + p_{11} p_{12} p_{22}) p_h H\\
EC_5 =& (p_{11}^4 + 3 p_{11}^2 p_{12} p_{21} + 2 p_{11} p_{12} p_{21} p_{22} + p_{12}^2 p_{21}^2 + p_{12} p_{21} p_{22}^2) p_l L && (6.A.6)\\
& + (p_{11}^3 p_{12} + p_{11}^2 p_{12} p_{22} + 2 p_{11} p_{12}^2 p_{21} + p_{11} p_{12} p_{22}^2 + 2 p_{12}^2 p_{21} p_{22} + p_{12} p_{22}^3) p_h H.
\end{aligned}
$$

Here, p_l and p_h are as defined above, while H and L are average health care cost of high (low) risks. For GR, premiums must reflect expected lifetime costs; e.g. the premium in period five equals expected health care cost in the first period ($P_5 = EC_1$; see PKH). This yields the following premium schedule,

$$
\begin{aligned}
P_1 =& p_{11}^3 p_l L && (6.A.7)\\
& + p_{12}([p_{11}^2 p_{21} + p_{11} p_{21} p_{22} + p_{12} p_{21}^2 + p_{21} p_{22}^2] p_l L + [p_{11} p_{12} p_{21} + 2 p_{12} p_{21} p_{22} + p_{22}^3] p_h H)\\
& + p_{11} p_{12}([p_{11} p_{21} + p_{21} p_{22}] p_l L + [p_{12} p_{21} + p_{22}^2] p_h H)\\
& + p_{11}^2 p_{12}(p_{21} p_l L + p_{22} p_h H) + p_{11}^3 p_{12}(p_h H - p_l L)\\
P_2 =& p_{11}^2 p_l L + p_{12}([p_{11} p_{21} + p_{21} p_{22}] p_l L + [p_{12} p_{21} + p_{22}^2] p_h H) && (6.A.8)\\
& + p_{11} p_{12}(p_{21} p_l L + p_{22} p_h H) + p_{11}^2 p_{12}(p_h H - p_l L)\\
P_3 =& p_{11} p_l L + p_{12}(p_{21} p_l L + p_{22} p_h H) + p_{11} p_{12}(p_h H - p_l L) && (6.A.9)\\
P_4 =& p_l L + p_{12}(p_h H - p_l L) && (6.A.10)\\
P_5 =& p_l L. && (6.A.11)
\end{aligned}
$$

These formulae yield the values presented in Table 6.1 of the text. The assumptions are $L = H$; $p_{11} = 1 - p_l$; $p_{12} = p_l$; $p_{13} = p_{24} = p_{21} = 0$; $p_{22} = 1$ (PKH; LHS of Table 6.1) and $L = H$; $p_{11} = 1 - p_l$; $p_{12} = p_l$; $p_{13} = p_{24} = 0$; $p_{21} = 1 - p_h$; $p_{22} = p_h$ respectively (Modified; RHS of Table 6.1). The parameter values are $p_l = 0.1$, $p_h = 0.25$, and $L = 20$.

As stated in the text GR premiums depend on weighted averages rather than the difference between expected cost pertaining to high and low risks. This is shown for the third period as follows. Adding $(+p_{12}p_l L - p_{12}p_l L)$ to the RHS of Equation (6.A.9) yields,

$$P_3 = p_{11}p_l L + p_{12}(p_{21}p_l L + p_{22}p_h H) + p_{11}p_{12}(p_h H - p_l L) + p_{12}p_l L - p_{12}p_l L. \quad (6.A.12)$$

Rearranging terms on the RHS, one obtains,

$$\begin{aligned} P_3 &= (p_{11} + p_{12})p_l L + p_{12}(p_{21}p_l L + p_{22}p_h H - p_l L) + p_{11}p_{12}(p_h H - p_l L) \\ &= p_l L + p_{12}(p_{21}p_l L + p_{22}p_h H - p_l L) + p_{11}p_{12}(p_h H - p_l L). \quad (6.A.13) \end{aligned}$$

The first term on the RHS of Equation (6.A.13) covers expected health care cost of a low risk during the current period. The second term covers the risk of an insured turning into a high risk in period two, the second to last period of the contract. The fact that she or he may become a low risk again in period three is accounted for by taking a weighted average of the cost pertaining to low and high risk respectively. The last term covers the risk of a person turning into a high risk in period three, in full analogy with PKH.

Appendix B

This appendix evaluates the effect of changing transition probabilities in the low-risk state and their corresponding effects on premiums P_t and P_{t+1} for the timing depicted in Figure 6.2 (c). The actuarially fair premium $P_{t+1} = (1+g)(p_l L + p_{13}C_l)$ and Equation (6.23c) with P_{t+1} replaced form the starting point,

$$\begin{aligned} P_{t+1} &= (1+g)(p_l L + p_{13}C_l) \quad (6.B.1) \\ P_t &= p_l L + p_{13}C_l + \quad (6.B.2) \\ &\quad \frac{1+g}{1+i}(p_{11}p_l L + p_{12}p_h H + p_{11}p_{13}C_l + p_{12}p_{24}C_h - (1-p_{13})(p_l L + p_{13}C_l)). \end{aligned}$$

The derivatives with respect to the transition probabilities are,

$$\frac{\partial P_t}{\partial p_{11}} = \frac{\partial p_{13}}{\partial p_{11}}C_l + RX + \frac{\partial p_{12}}{\partial p_{11}}RY + \frac{\partial p_{13}}{\partial p_{11}}RZ \quad (6.B.3)$$

$$\frac{\partial P_{t+1}}{\partial p_{11}} = \frac{\partial p_{13}}{\partial p_{11}}(1+g)C_l \quad (6.B.4)$$

$$\frac{\partial P_t}{\partial p_{12}} = \frac{\partial p_{13}}{\partial p_{12}}C_l + \frac{\partial p_{11}}{\partial p_{12}}RX + RY + \frac{\partial p_{13}}{\partial p_{12}}RZ \quad (6.B.5)$$

$$\frac{\partial P_{t+1}}{\partial p_{12}} = \frac{\partial p_{13}}{\partial p_{12}}(1+g)C_l \quad (6.B.6)$$

$$\frac{\partial P_t}{\partial p_{13}} = C_l + \frac{\partial p_{11}}{\partial p_{13}}RX + \frac{\partial p_{12}}{\partial p_{13}}RY + RZ \quad (6.B.7)$$

$$\frac{\partial P_{t+1}}{\partial p_{13}} = (1+g)C_l \quad (6.B.8)$$

where R, X, Y, and Z are defined as follows,

$$R = (1+g)/(1+i) \quad (6.B.9)$$

$$X = p_l L + p_{13}C_l \quad (6.B.10)$$

$$Y = p_h H + p_{24}C_h \quad (6.B.11)$$

$$Z = p_l L + p_{13}C_l - p_{12}C_l. \quad (6.B.12)$$

Without additional assumptions regarding the remaining parameter values and the effect on composite probabilities [three scenarios in analogy of the scenarios defined below Equation (6.12) can be deduced] no predictions can be derived.

Appendix C

The calculation of net gain in expected utility due to GR health insurance is considered here. As in Frick (1998), a logarithmic risk utility function is assumed. For the utility gained from insurance ($\triangle EU$), in the first period one is left with,

$$\begin{aligned} \triangle\ EU_1 &= (1-p_{13})\{\ln[p_l(y-L)+(1-p_l)y] - p_l \ln[y-L] - (1-p_l)\ln[y]\} \\ &= 0.0020. \end{aligned} \tag{6.C.1}$$

With cost growth g and interest i, one obtains for the second period,

$$\begin{aligned} \triangle\ EU_2 &= \{(1-p_{13})p_{11} + (1-p_{24})p_{12}\} \\ &\quad \{\ln[(p_{11}p_l + p_{12}p_h)(y-L(1+g)) + (p_{11}(1-p_l)+p_{12}(1-p_h))y]\} \\ &\quad -\{(1-p_{13})p_{11} + (1-p_{24})p_{12}\}\{(p_{11}p_l + p_{12}p_h)\ln[y-L(1+g)]\} \\ &\quad -\{(1-p_{13})p_{11} + (1-p_{24})p_{12}\}\{(p_{11}(1-p_l)+p_{12}(1-p_h))\ln[y]\} \\ &= 0.1027 \end{aligned} \tag{6.C.2}$$

EU_2 is much larger than EU_1 because every insured is a low risk in the first period, whereas the probability of loss is much higher in the next period. In total, the utility premium obtained is,

$$\begin{aligned} \triangle\ EU &= \triangle\ EU_1 + \beta(1-p_{13})\triangle\ EU_2 \\ &= 0.0020 + 0.0996\beta > 0 \end{aligned} \tag{6.C.3}$$

The loss in expected utility caused by the binding restriction $P_2 = p_l L$ can be calculated by determining the optimal first-period premium P_1^* when the Lagrange-multiplier is $\mu = 0$ in the optimization problem in Equation (6.20). This value is used to calculate the increase in expected utility ($\triangle\ EU_1^* > 0$) in period 1 which is the difference between the optimized utility and the utility with a binding restriction,

$$P_1^* = \frac{106.7\beta - 92.1211}{1.1 + 1.067\beta}; \tag{6.C.4}$$

$$\triangle\ EU_1^* = \ln\left[100 - \frac{106.7\beta - 92.1211}{1.1+1.067\beta}\right] - 4.5797. \tag{6.C.5}$$

However, in period 2 the binding restriction causes a reduction in expected utility ($\triangle\ EU_2^* < 0$) compared to the optimized value of P_2^*. The reduction is given by,

$$P_2^* = 100 - 110\beta + \frac{1.1\beta(106.7\beta - 92.1211)}{1.1 + 1.067\beta}; \quad (6.C.6)$$

$$\triangle\ EU_2^* = \ln\left[110\beta - \frac{1.1\beta(106.7\beta - 92.1211)}{1.1 + 1.067\beta}\right] - 4.5812. \quad (6.C.7)$$

The total reduction of expected utility from the binding constraint therefore amounts to,

$$\triangle\ EU^* = \underbrace{\triangle\ EU_1^*}_{>0} + \beta(1 - p_{13}) \underbrace{\triangle\ EU_2^*}_{<0} > 0 \quad (6.C.8)$$

The difference between (C.3) and (C.8) is the net gain in expected utility from having GR insurance.

$$\triangle_n\ EU = \triangle\ EU - \triangle\ EU^* \quad (6.C.9)$$

With the parameter values taken from Table 6.2 the critical value for $\triangle_n\ EU = 0$ of $\hat{\beta}$ is 0.526. Figure 6.5 illustrates the difference as a function of β.

Bibliography

Beck, K. and P. Zweifel (1998). Cream-skimming in deregulated social health insurance: Evidence from Switzerland. In P. Zweifel (Ed.), *Health, the Medical Profession and Regulation – Developments in Health Economics and Public Policy, Volume 6*, pp. 211–227. Dordrecht: Kluwer.

Bhattacharyya, G. and R. Johnson (1977). *Statistical Concepts and Methods.* New York: J. Wiley & Sons.

Brown III, H. S. and L. B. Conelly (2005a). Lifetime cover in private insurance markets. *International Journal of Health Care Finance and Economics 5*(1), 75–88.

Brown III, H. S. and L. B. Conelly (2005b). Market failure in long-term private health insurance markets: A proposed solution. *Applied Economics Letters 12*(5), 281–284.

Cochrane, J. H. (1995). Time-consistent health insurance. *The Journal of Political Economy 103*(3), 445–473.

Federal Statistical Office (BFS) (2007). *Mortality tables of Switzerland 1998/2003*. Neuchatel: Federal Statistical Office (BFS).

Frick, K. D. (1998). Consumer capital market constraints and guaranteed renewable insurance. *Journal of Risk and Uncertainty 16*(3), 271–278.

Herring, B. and M. V. Pauly (2001). Premium variation in the individual health insurance market. *International Journal of Health Care Finance and Economics 1*(1), 43–58.

Herring, B. and M. V. Pauly (2006). Incentive-compatible guaranteed renewable health insurance premiums. *Journal of Health Economics 25*(3), 395–417.

Lubitz, J. D. and G. F. Riley (1993). Trends in Medicare payments in the last year of life. *The New England Journal of Medicine 328*(15), 1092–1096.

Nickel, A. (2005). Versicherung der Veränderung der Risikosituation in der PKV - (Insuring a change in risk in German private health insurance). *Zeitschrift für Versicherungswesen* (16), 505–509.

Patel, V. and M. V. Pauly (2002). Guaranteed renewability and the problem of risk variation in individual health insurance markets. *Health Affairs 21*(5), 280–289.

Pauly, M. V., H. Kunreuther, and R. Hirth (1995). Guaranteed renewability in insurance. *Journal of Risk and Uncertainty 10*(2), 143–156.

Pauly, M. V., A. Nickel, and H. Kunreuther (1998). Guaranteed renewability with group insurance. *Journal of Risk and Uncertainty 16*(3), 211–221.

Steinmann, L., H. Telser, and P. Zweifel (2007). Aging and future healthcare expenditure: A consistent approach. *Forum for Health Economics & Policy 10*(2), Article 1, http://www.bepress.com/fhep/10/2/1.

Werblow, A., S. Felder, and P. Zweifel (2007). Population ageing and health care expenditure: A school of 'red herrings'? *Health Economics 16*, 1109–1126.

Zweifel, P., S. Felder, and M. Meier (1999). Ageing of population and health care expenditure: A red herring? *Health Economics 8*(6), 485–496.

Chapter 7

Conclusion

The conclusion is first confined to a short discussion of policy implications, and then proceeds to possible extensions and improvements with regard to the unpublished chapters 4, 5, and 6.

Chapters 2 and 3 clearly imply that policy makers need to pay attention when increasing health care expenditure or social security in general. The findings disclose that in both cases additional expenditure is created through a feedback process. Chapter 4's analysis suggests that positive short-term correlations increase the variance in consumer's assets and policy makers should pay attention to such positive correlations. However, chapter 4's empirical estimations also reveal the absence of long-term common trends. Domestic and international hedging of private lines of insurance seems possible, thus enabling insurers to offer combined contracts. This calls for reducing any barriers to trade. Chapter 5 shows that when premiums are community-rated any cap on the volume of risk adjustment induces additional risk-selection efforts by competing health insurers. When policy makers consider an additional criterion in the formula of risk-adjustment, and a cap at the same time, they must evaluate each additional criterion and its impact on risk selection again. Admittedly, a more efficient way to reduce volume would be to allow insurers to set premiums that vary with some risk factors. Finally, the model in chapter 6 illustrates that imposing a guaranteed renewability (GR) mandate would result in an overall efficiency loss. Indeed, a large fraction of the population would voluntarily participate in GR without such a mandate.

Various improvements and extensions can be identified with regard to chapters 4 to 6. A great improvement to chapter 4 would be to use individual level data. Possible aggregation of individual level data to firm levels would allow using copulas to gain in depth understanding of (nonlinear) dependence structures between different lines of insurance at the firm level [see Embrechts et al. (2003)]. Ideally this data could be combined with information on wealth, health, and wisdom [this could be done with the new anonymous Swiss AHV/AVS number and data from health insurers, see Widmer et al. (2007) for a similar idea regarding premium subsidies in health insurance]. Testing the theory using individual level data would dissolve the problem of possible aggregation bias. Individual companies could estimate their combined risk across several lines of insurance within and across countries using copulas in addition to the analysis of short-term correlations and long-term common trends. Analyzing short-term linear correlations define a legitimate approach if the hypothesis of a multivariate normal distribution (Gauss-Copula) cannot be rejected. Combining the results of both econometric methods seems like a valuable path for future research.

One can think of many extensions to chapter 5. Arguably, the discussion must revolve around improving the formula of risk adjustment. Adding additional criteria to the formula will most likely affect its volume – an effect that nonetheless can be assessed using empirical methods. The additional criteria that could be included have been subject to an extensive discussion and are cited in section 5.3. Changing the formula of risk adjustment when premiums are community-rated will affect risk selection between competing health insurers. This observation defines the starting point in order to advance the theory presented in chapter 5. The reduction in variance when including an additional criterion in the formula of risk adjustment and the increase due to an imposed cap should simultaneously be combined in one optimization. Optimizing and calculating the effect of several additional criteria should then illustrate which additional criterion induces the least risk-selection efforts in combination with a cap. In addition, different measures of risk could replace the chosen objective function. Either value at risk, or expected loss at risk, would disclose whether the (uniform) reductions per risk class derived in chapter 5 are still optimal.

Chapter 6 develops a theoretical model and derives results that could easily be tested with individual level data from a health insurer. One future research question would be to determine the relationship between cost of death for low and high risks. This can be accomplished both when data on specific risks is available [as in Herring and Pauly (2006)], or when data on health care expenditure is available to derive health status from past expenditure [as in

Lehmann and Zweifel (2004)]. Those GR premiums that also cover the cost of dying can therefore be split into two parts, and estimation based on individual level data could derive life-time GR premiums. Furthermore, the model could be adapted for use in life insurance, or for housing insurance in flood and earthquake areas (the death state could denote complete destruction of housing). In fact, the model is usefully adopted to almost any kind of insurance contracts. A further theoretical enhancement of the process depicted in chapter 6 could be to compare community-rated and risk-rated GR premiums. The community-rated premium schedule mirrors expected cost, while the risk-rated GR premium reverses it. Fixing the duration of the contract and the subjective discount rate of a consumer both contract's utility could be derived and compared. With everyone starting in low-risk status as in chapter 6, very impatient individuals should prefer community-rated premiums. In addition, chapter 6 also points to the importance of chronology. Calculating an accurate GR premium schedule involves determining whether the premium is paid before or after transition. Extending research to include chronologies in different countries or states would shed light on the regional value of critical subjective time preference when death is included. More extensions with regard to alternative utility functions, modelling age dependent loss probabilities, or supplementary risk states will result in enlightening additional conclusions.

Bibliography

Embrechts, P., F. Lindskog, and A. McNeil (2003). Modelling dependence with copulas and applications to risk management. In S. Rachev (Ed.), *Handbook of Heavy Tailed Distributions in Finance*, pp. 329–384. Amsterdam: Elsevier.

Herring, B. and M. V. Pauly (2006). Incentive-compatible guaranteed renewable health insurance premiums. *Journal of Health Economics 25*(3), 395–417.

Lehmann, H.-J. and P. Zweifel (2004). Innovation and risk selection in deregulated social health insurance. *Journal of Health Economics 23*, 997–1012.

Widmer, W., K. Beck, L. Boos, L. Steinmann, and R. Zehnder (2007). *Eigenverantwortung, Wettbewerb und Solidarität - (Personal responsibility, competition, and solidarity)*. Zurich: Schriftenreihe der SGGP.

Curriculum Vitae

Patrick Eugster was born on the 8th of January 1977 in Kilchberg, Zurich (Switzerland) and attended primary school in Feldmeilen, Zurich. After high school at the "Kantonsschule Hohe Promenade" in Zurich, various exchange years in the United States of America (High School), Italy and England (both during University) he completed his university studies focusing on microeconomics and management at the University of Zurich in December 2003. Thereafter he began his doctoral thesis at the chair of Prof. Dr. P. Zweifel (Chair for Applied Microeconomics focused on health-, insurance-, and energy-economics as well as international trade). During his time at the chair of Prof. Dr. P. Zweifel he supervised various university courses and carried out several projects in collaboration with Swiss health insurers.