Université catholique de Louvain
Institute of Statistics, Biostatistics and Actuarial Sciences

Actuarial models for health and long-term care insurance

—

Nathalie LUCAS

A thesis submitted to the
Université catholique de Louvain
in partial fulfillment of the
requirements for the degree of
DOCTOR OF SCIENCES

Thesis Committee:

Prof. **Michel DENUIT**	*Université catholique de Louvain*	Advisor
Prof. **Pierre DEVOLDER**	*Université catholique de Louvain*	Chairman
Prof. **Jan DHAENE**	*KU Leuven*	
Prof. **Donatien HAINAUT**	*Université catholique de Louvain*	
Prof. **Peter HIEBER**	*University of Ulm*	

April 2021

Ce n'est pas parce que les choses sont difficiles que nous n'osons pas, c'est parce que nous n'osons pas qu'elles sont difficiles.

Sénèque

Acknowledgements

Dans une lettre à son fils, Einstein a écrit ‘la vie c’est comme une bicyclette, il faut avancer pour ne pas perdre l’équilibre’. Si j’ai pu avancer continûment dans cette aventure, c’est sans conteste grâce au soutien de différentes personnes.

Tout d’abord, je suis très reconnaissante envers mon promoteur de thèse, monsieur le Professeur Michel Denuit. Je le remercie pour sa confiance ; la gigantesque expertise technique qu’il a partagée ; sa grande disponibilité et, non des moindres, son empathie et ses qualités humaines d’écoute et de compréhension. Ces éléments ont été déterminants dans la réussite de mon aventure doctorale.

Je souhaiterais exprimer ma gratitude à messieurs les Professeurs Pierre Devolder et Jan Dhaene pour avoir accepté de faire partie de mon comité d’encadrement. Les différentes remarques, notamment lors de l’épreuve de confirmation ou lors de la défense privée, ont contribué à l’amélioration significative de mes travaux.

Je remercie particulièrement messieurs les Professeurs Donatien Hainaut et Peter Hieber pour avoir participé à ce jury de thèse et contribué par là à l’amélioration de mes travaux.

Je suis également reconnaissante envers les personnes impliquées dans l’élaboration de mes différents travaux: pour leur collaboration et leur disponibilité, je remercie Hervé Avalosse, Marcus Christiansen, Hamza Hanbali, Peter Hieber, Ermano Pitacco, Jan-Philipp Schmidt et Julien Trufin.

Je remercie la chaire DKV pour son support financier durant mes 2 premières années de doctorat.

Mon travail de recherche au quotidien a été grandement facilité par un environment de travail chaleureux, si particulier à l’ISBA. Je pense au soutien administratif, logistique et calorifique (avec la ’pause biscuits’) du staff composé de Maguy, Nadja, Nancy, Sophie et Tatiana. Merci pour votre super organisation, votre disponibilité et votre bonne humeur.

Bien entendu je pense à tous mes collègues et anciens collègues de bureau avec qui j’ai partagé des moments mémorables. Un clin d’oeil à Alexandre, Anna, Antoine, Aurélie, Benjamin, Charles, Emmanuel, Fadoua, Florian, Gilles, Hélène, John-John, Lexuri, Mailis, Mickael, Nathan, Pauline, Rebecca, Sophie, Stefka, Sylvie, Vincent B. et Vincent P.

Je remercie l’ensemble des professeurs de l’ISBA, pour la bonne ambiance et la qualité

de l'environment de savoir et de transmission de savoir de l'institut.

Un merci particulier à l'équipe informatique (Christophe, Mickael, Pierre, Raphael) pour leur bonne humeur et leur extrême gentillesse.

Je remercie le SMCS pour ses missions de consultance.

Enfin je souhaite dédier ce travail à tous mes proches et amis, notamment tous mes amis de mon club de tennis de la raquette de Wavre, ainsi que ceux de mon club de trail de LLN, le James. Les multiples moments de détente sportive et amicale ont été vitaux.

Un dernier mot pour ma famille, en particulier mes parents ainsi que mes frères: je vous remercie du fond du coeur de toujours croire en moi.

Contents

I Introduction **1**

1 Introduction **3**
1.1 Overview of the health insurance market . . . 3
1.2 Objectives . . . 4
1.3 Outline . . . 5
1.3.1 Part II: Hospitalization insurance . . . 5
1.3.2 Part III: Long-term Care insurance . . . 6

II Hospitalization insurance **11**

2 Premium and provision adjustment to trends **13**
2.1 Introduction . . . 13
2.2 Actuarial model . . . 14
2.2.1 Two-decrement model . . . 14
2.2.2 Benefits and level premiums . . . 15
2.3 Adapting the premium and/or the reserve level at time 1 . . . 16
2.3.1 Accumulated reserve . . . 16
2.3.2 Revision of benefits . . . 16
2.3.3 Premium and/or reserve update . . . 17
2.3.4 Adapting the premium, only . . . 18
2.4 Adapting the premium level at time k . . . 19
2.4.1 Accumulated reserve . . . 19
2.4.2 Premium update . . . 20
2.4.3 Adaptation to age-specific inflation . . . 22
2.5 Case study: The new indexing mechanism for the Belgian medical insurance market . . . 23
2.5.1 Indexing rule imposed by the Belgian law . . . 23

2.5.2 Technical basis . . . 24
2.5.3 Effect of inflation . . . 24
2.6 Future perspectives . . . 29
2.6.1 Risk transfer mechanism . . . 29
2.6.2 Risk factors . . . 32
2.7 Conclusion . . . 33

3 Stochastic modelization of claim costs trends 35
3.1 Introduction . . . 35
3.2 Bilinear modeling . . . 36
3.2.1 Model specification . . . 36
3.2.2 Identifiability . . . 36
3.2.3 Estimation . . . 37
3.2.4 Forecast . . . 37
3.3 Case study . . . 37
3.3.1 Data description . . . 37
3.3.2 The Rusam method . . . 39
3.3.3 Modeling approach . . . 40
3.3.4 Model comparisons . . . 40
3.3.5 Model selection . . . 41
3.3.6 Projection . . . 46
3.4 Remark on overfitting . . . 50
3.5 Conclusion . . . 52

4 Joint analysis of mortality and morbidity trends 57
4.1 Introduction . . . 57
4.2 Modeling strategy . . . 59
4.2.1 Data under consideration . . . 59
4.2.2 Frequency-severity decomposition . . . 59
4.2.3 Nonlinear regression models . . . 60
4.2.4 Projection . . . 60
4.3 Numerical illustration . . . 61
4.3.1 Data . . . 61
4.3.2 Descriptive statistics . . . 62
4.3.3 Generalized Additive Modeling . . . 65
4.3.4 Comparison with an aggregate loss model . . . 69
4.3.5 Out-of-sample analysis . . . 71
4.3.6 Projections . . . 71
4.4 Conclusion . . . 77

5 Part II: conclusion 81

III Long-term care insurance 83

6 Insurance approach 85
6.1 Introduction . . . 85
6.2 Multistate modeling . . . 86
6.3 Transition probabilities . . . 87
6.4 Transition intensities . . . 87
6.5 Equivalence principle . . . 89
6.6 Generalized annuities . . . 92
6.7 Generalized life insurances . . . 93
6.8 Some specific conditions . . . 94
6.8.1 Insured period . . . 94
6.8.2 Waiting, or elimination period . . . 94
6.8.3 Deferred period . . . 95
6.9 Premium formulas for some LTC insurance products . . . 95
6.9.1 Stand-alone LTC cover . . . 96
6.9.2 Enhanced pension, or life care annuity . . . 97
6.9.3 Package of LTC and lifetime-related benefits . . . 99
6.9.4 Whole-life insurance with LTC acceleration benefit . . . 100
6.9.5 LTC package with a whole-life insurance offsetting the deferred period . . . 102
6.10 Reserves . . . 103
6.10.1 Principle . . . 103
6.10.2 Reserve formulas for some LTC insurance products . . . 107
6.11 Conclusion . . . 109

7 Collaborative approach or P2P 113
7.1 Introduction . . . 113
7.2 2-state framework . . . 114
7.2.1 Tontine payoff . . . 115
7.2.2 Actuarial fairness . . . 118
7.2.3 Numerical example 1 . . . 120
7.3 3-state framework . . . 123
7.3.1 Additional notation . . . 123
7.3.2 Life-care annuity . . . 124
7.3.3 Life-care tontine . . . 124
7.4 Conclusion . . . 136

IV Discussion and Extensions 141

8 Discussion and extensions 143

Part I

Introduction

Chapter 1

Introduction

1.1 Overview of the health insurance market

In the CEA-Groupe Consultatif "Solvency II Glossary", health insurance is considered as a "generic term applying to all types of insurance indemnifying or reimbursing losses (e.g. loss of income) caused by illness or disability, or for expenses of medical treatment necessitated by illness or disability". Its market share is shared between private providers and public providers.

In Belgium, public health insurance is organized by the state through the Federal Agency RIZIV-INAMI and operated by several so-called "sickness funds" (non-profit organizations). As pointed out by Schokkaert and Van de Voorde (2003), a few large sickness funds dominate the Belgian market of compulsory health insurance. The entire population is covered and benefits from a very broad package, including e.g. ambulatory and dental care. While membership of a sickness fund is compulsory, every individual can enroll in the sickness fund of his or her choice. Sickness funds historically developed along political and religious lines and are grouped at the national level in five associations. The two most important ones, the Christian Mutualities and the Socialist Mutualities, insure together about 75% of the population. In addition there is one public fund mainly acting as a kind of "insurer of last resort".

Besides the broad benefit package comprised in the compulsory cover, individual or group private health insurance contracts are sold which pay (part of) the non-covered medical costs and supplements, for which a separate premium is charged. These private insurance products, offering additional optional insurance, are regulated via the so-called "Law Verwilghen" of 20 July 2007 and "Law Reynders" (also called the "Law Verwilghen II") of 17 June 2009, both named after the ministers in charge. One notable characteristic is their lifelong nature (see Law Verwilghen).

The analysis performed in this thesis is of interest for both sickness funds and private insurers selling optional coverages.

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Male	76.9	77.4	77.5	78	77.8	78.1	78.8	78.7	79	79.2	79.4
Female	82.6	82.8	83	83.3	83.1	83.2	83.9	83.4	84	83.9	83.9

Table 1.1: Belgian life expectancy at birth (source: Eurostat)

2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
25.8	25.9	26	26	26.5	26.9	27.4	27.8	28.2	28.6	29.1	29.5

Table 1.2: Belgian old-age dependency ratio (source: Eurostat)

1.2 Objectives

Actuaries working in life and pension insurance have been using projected life tables for several decades. However, the mortality improvements seen in practice have quite consistently exceeded the projected improvements. From the actuarial risk management perspective, the major problem is that mortality improvement is not a diversifiable risk. Traditional diversifiable mortality risk is the random variation around a fixed, known life table. Mortality improvement risk, though, affects the whole portfolio and thus can not be managed using the law of large numbers. The same comments for mortality apply to morbidity in health insurance and it is even more tricky to forecast, as it is impacted by different drivers including longevity and medical inflation. Modifications in law or in regulation also affect more rapidly morbidity and health expenses than mortality.

One forthcoming challenge for the health insurance industry is definitely related to population aging and the increasing life expectancy (see Table 1.1). Demographic projections performed by Eurostat show that profound changes are expected in the population age structure. This results notably from a low fertility and an increasing life expectancy, resulting in an important old-age population. The increasing old-age dependency ratio (see Table 1.2 for Belgium), defined as the ratio of the number of elderly people (i.e. aged 65 and over) compared to the number of people of working age, is general in the European states.

Therefore some new products targeting the particular needs of an aging population are crucially required. Specific trends linked to morbidity and mortality need to be analysed and anticipated. In this thesis we propose to analyse how systematic risks, especially morbidity and inflation, can be efficiently managed, exploiting the correlation structure of the risks and some innovative risk sharing agreements with policyholders. The analysis uses multistate models for disability and health insurance integrating longevity and inflation. see Christiansen et al. (2012).

Therefore we specifically consider health insurance products targeting an aged, i.e. post-retirement, population. We looked into two products: hospitalization insurance and long-term care (LTC) insurance, which are the focus of parts II and III, respectively. Since

these products are long-term, it is essential to examine carefully possible trends in the transition rates and in the claims, so as to determine appropriate pricing and reserving rules.

Please note that the term "hospitalization insurance" was chosen as it is the term adopted by the Belgian Financial Services and Market Authority (FSMA), see FSMA (2020) for more details on the products sold in Belgium.

The two products may differ in several aspects.

- type of benefits: benefits can be indemnitary or reimbursement, which is more common in the hospitalization cover ; or a lump sum or predefined benefits, which is more common in the LTC cover.

- type of premium: coverage is lifelong but annual premiums are more common in the hospitalization cover. A single premium is frequent in LTC, inducing a big uncertainty on the technical basis or a big systematic pricing risk at initiation.

- risk-sharing mechanism: systematic or non-diversifiable risk can be shared between the policyholder and the insurer. The risk can be borne by the policyholder by adapting the premium or the benefit, or by the insurer by adapting the reserve. A collaborative approach is an alternative proposed in chapter 6.

The two products will be treated separately, in parts II and III.

1.3 Outline

1.3.1 Part II: Hospitalization insurance

In the Belgian law dealing with private hospitalization insurance (Verwilghen 20/7/2007), annual premiums are allowed to be revised annually. Indeed in view of the observed changes in longevity, morbidity and economic conditions, it appears extremely risky to specify insurance benefits in absolute terms. The health benefits that will be paid over the years for a lifelong health insurance policy are impacted by unpredictable changes in prices for medical goods and services. Given the long-term nature of health insurance contracts and the impossibility to predict or hedge against medical inflation, insurers are not able to appropriately account for this medical inflation in the calculation of the yearly premium level at policy issue. Therefore, these lifelong contracts are usually designed in such a way that the insurer is allowed to adapt the premium amounts at regular times (e.g. yearly) to account for medical inflation not taken into account at policy issue, based on some predefined medical inflation index. This practice is used in several EU member countries. Lifelong health insurance contracts and related premium updating mechanisms have been investigated in Vercruysse et al. (2013) and is the main focus of chapter 2.

Chapter 2 introduces indeed an a-posteriori premium adjustment to take into account inflation, and potentially other systematic risks. We show that ex-post indexing can be

achieved by considering only premiums, without explicit reference to reserves. This appears to be relevant in practice as reserving mechanisms may not be transparent to policyholders and as some insurers do not compute contract-specific reserves, managing the whole portfolio in a collective way.

Chapter 3 and 4 propose stochastic models to analyze inpatients claims evolution. This part of the thesis presents an analytical framework for understanding the relationship between mortality and morbidity and how this relationship might evolve over time. In chapter 3 costs are modelized in a Lee-Carter type regression, i.e. particular case of bivariate function, using private insurer data. This chapter proposes a practical way for modeling and projecting health insurance (inpatients) expenditures over short time horizons, based on observed historical data. It is motivated by a similar age structure generally observed for health insurance claim frequencies and yearly aggregate losses on the one hand and mortality on the other hand. As an application, the approach is illustrated for German historical inpatient costs provided by the Federal Financial Supervisory Authority (BaFin).

In addition to the age and time variables introduced in chapter 3, chapter 4 studies the dynamics in end-of-life inpatients hospital expenses. Miller (2001) first states that integrating proximity-to-death greatly modifies forecasts. The literature suggests a direct association between high expenditures and death: for example in the American Medicare program, the 5 percent of beneficiaries aged 65 and over who die each year account for 25 to 30 percent of total expenditures (Hoover et al. 2002). We propose a refinement of the dual approach decedents, i.e. dying within the year, vs survivors, i.e. dying after one year (Yang et al. 2003). The proposed model is based on a frequency-severity decomposition including age, calendar time, longevity dynamics, and time-to-death. These features are treated as continuous explanatory variables in nonlinear regression models with Poisson, Gamma and Tweedie error structures. Proximity to death is controlled for, as well as longevity improvements by including projected life tables into the proposed model. This allows the analyst to isolate the different effects impacting late-life hospital costs. A detailed case study is performed on Belgian data and produces projections over short to medium time horizons. We show that total costs are mainly driven by the frequency component for the data under consideration.

1.3.2 Part III: Long-term Care insurance

This part relates to a specific health insurance product: Long-Term Care (LTC) Insurance. The LTC insurance policies, which concern millions of individuals, are at present very heterogeneous, using many types of guarantees and many types of benefits underwriting modes. Markets are offering since the beginning payments of monthly lifetime cash annuities. Yet the growing LTC market is currently proposing some indemnity-based products. In brief, benefits in LTC insurance products can be classified in 3 main categories:

- Predefined benefits

  Benefits of a predefined amount (usually, a lifelong annuity benefit, for example on a monthly or quarterly basis) can be either a fixed-amount benefit, stated in the policy conditions, or a degree-related (or graded) benefit. A graded benefit is a benefit

whose amount is graded according to the degree of disability, that is, the severity of the disability itself; the severity must be assessed relying on a scoring system, for example the Activities of Daily Living (ADL) scale. Benefits of a predefined amount can be provided by a stand-alone LTC cover as well as by several types of combined products.

- Reimbursement benefits

  This category includes LTC insurance products which provide expense reimbursement. Two basic types of products can be recognized. Stand-alone LTC cover provides the (partial) reimbursement of expenses related to LTC needs, in particular nursery, medical expenses, physiotherapy, etc. Usually, there are limitations on eligible expenses. Further, deductibles (in terms of fixed amount or fixed percentage, or a combination of both) as well as limit values are stated in the policy conditions. LTC benefits can also be provided by an LTC cover as a rider to sickness insurance. The resulting product is a lifelong sickness insurance. However, in order to cover LTC needs, eligible expenses are extended with respect to a usual sickness cover, so to include, for example, nursing home expenses. Further, a fixed-amount daily benefit can be paid to cover expenses without documentary evidence.

- Service benefits

  The LTC insurance products providing care service benefits usually rely on an agreement between an insurance company and an institution which acts as the care provider. An interesting alternative is given by the Continuing Care Retirement Communities, briefly CCRCs, which have become established in the US. CCRCs offer housing and a range of other services, including long-term care. Costs (in particular related to LTC) are usually met by a combination of entrance charge plus periodic fees (that is, upfront premium plus monthly premiums).

In part III we focus on predefined benefits.

LTC insurance is currently experiencing an increasing demand. Triggered by population aging, LTC costs have shown a significant increase over the recent decades. In the US, data by the National Health Expenditures Account (NHEA) show that expenditures in the Medicaire program, aiming to support US residents with low income in long-term care, raised from $225 billion in 2000 (2.2% of the gross domestic product (GDP)) to $750 billion in 2018 (3.6% of GDP). Also governmental spending in home health care raised from $32 billion in 2000 to $102 billion in 2018. A similar observation is made in Europe, for instance in Belgium, where LTC spending (in terms of GDP) increased from 1.7% in 2000 to 2.3% in 2018 (source: Eurostat).

The increasing trend of LTC costs is projected to continue in the future (see Shi 2013). According to United Nations projections, the number of elderly people, i.e. older than 65, is projected to triple from 2020 to 2080 to reach 2.2 billion. The global share of the elderly population is expected to rise from 9.4% in 2020 to 20.6% in 2080, and the demand for long-term care services in the years to come is expected to further increase.

Specific insurance products are dealing with LTC risk, notably the classical LTC cover, which provides benefits in case of dependency, and the enhanced pension or life-care annuity. The latter combines regular payments of a life annuity with LTC insurance. In terms of risk management, the pooling of competing risks, i.e. longevity and morbidity, is quite advantageous as the two risks act in opposite directions (see Murtaugh 2001). Indeed given the correlation between illness and mortality demands for annuities and long-term care insurance interact. When moving into dependency, individuals receive higher benefits, but also suffer from a decrease in their life expectancy, creating a natural hedge.

The key advantages of the life-care annuity relative to the stand-alone products life annuity and classical LTC cover are its potential to decrease the costs and to make coverage available to more potential purchasers (see Spillman 2003). One reason for this is a reduction in adverse selection. In a life annuity, individuals with low longevity expectations are less likely to buy annuities, forcing insurance providers to increase their premiums accordingly. Indeed, it has been estimated that around 10% of the cost of life annuity premiums is due to adverse selection (see Friedman 1990). On the other hand, classical LTC covers are not available to everyone as underwriting mostly rejects people in bad health. Combining both products makes insurance affordable for people in a poor health state for whom it is currently unattractive to buy a life annuity and unaffordable to buy a classical LTC cover. Therefore a life-care annuity allows the inclusion of the currently rejected population, which lowers the cost for all and reduces adverse selection (see Brown 2013).

In part III we build on the advantage of pooling mortality and morbidity risks.

Chapter 6 provides pricing and reserving formulas for LTC related products including combined products, in a multi-state predefined benefits framework. Numerical illustrations are provided, based on French private market data. We use a 3-state, semi-Markov framework and analytical expressions have been obtained for the premiums and reserves of different LTC products, including combined products.

Chapter 7 proposes an innovative risk-sharing mechanism through a collaborative or P2P approach. We focus on classical mutual risk pooling schemes, i.e. tontines, and introduce a "life-care tontine", which in addition to retirement income targets the needs of long-term care coverage for an ageing population.

This scheme is actuarially fair at each time. Pooling heterogeneous risks (i.e. different age groups) is shown to reduce overall risk. The life-care tontine is compared to a classical life-care annuity. The model is applied to real life data, illustrating the adequacy of the proposed tontine scheme.

References

- Brown, J. & Warshawsky, M. (2013). The Life Care Annuity: A New Empirical Examination of an Insurance Innovation that Addresses Problems in the Markets for Life Annuities and Long-Term Care Insurance. The Journal of Risk and Insurance 80, 677-704.

- Christiansen, M., Denuit, M. & Lazar, D. (2012). The Solvency II square-root formula for systematic biometric risk. Insurance: Mathematics and Economics 50, 257-265.

- FSMA (2020). Hospitalisation Insurance.
Available from https://www.fsma.be/en/hospitalisation-insurance

- Friedman, B. & Warshawsky, M. (1990). The cost of annuities: Implications for saving behavior and bequests. The Quarterly Journal of Economics 105(1), 135-154.

- Hoover, D., Crystal, S., Kumar, R., Sambamoorthi, U. & Cantor, J. (2002). Medical expenditures during the last year of life: findings from the 1992-1996 Medicare current beneficiary survey. Health services research 37(6), 1625-1642.

- Miller, T. (2001). Increasing longevity and Medicare expenditures. Demography 38, 215-226.

- Murtaugh, C., Spillman, B. & Warshawsky, M. (2001). In sickness and in health: An annuity approach to financing long-term care and retirement income. Journal of Risk and Insurance, 225-253.

- Payne, G., Laporte, A., Deber, R. & Coyte, P. (2007). Counting backward to health care's future: using time-to-death modeling to identify changes in end-of-life morbidity and the impact of aging on health care expenditures. The Milbank Quarterly 85(2), 213-257.

- Schokkaert, E. & Van de Voorde, C. (2003). Belgium: Risk adjustment and financial responsibility in a centralised system. Health Policy 65, 5-19.

- Shi, P. & Zhang, W. (2013). Managed care and health care utilization: Specification of bivariate models using copulas. North American Actuarial Journal 17(4), 306-324.

- Spillman, B., Murtaugh, C. & Warshawsky, M. (2003). Policy Implications of an Annuity Approach to Integrating Long-Term Care Financing and Retirement Income. Journal of Aging and Health 15(1), 45-73.

- Vercruysse, W., Dhaene, J., Denuit, M., Pitacco, E. & Antonio, K. (2013). Premium indexing in lifelong health insurance. Far East Journal of Mathematical Sciences, 365-384.

- Yang, Z., Norton, E. & Stearns, S. (2003). Longevity and health care expenditures: the real reasons older people spend more. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 58(1), S2-S10.

Part II

Hospitalization insurance

Chapter 2

Premium and provision adjustment to trends

The present chapter is based on the following published papers:

- Denuit, M., Dhaene, J., Hanbali, H., Lucas, N., Trufin, J. (2017). Updating mechanism for lifelong insurance contracts subject to medical inflation. European Actuarial Journal 7, 133-163.

- Denuit, M., Dhaene, J., Hanbali, H., Lucas, N., Trufin, J. (2017). Hospitalisation: le nouveau méchanisme belge d'indexation. Monde de l'Assurance 2017.05, 39-46.

2.1 Introduction

Consider a portfolio of lifelong health insurance contracts covering medical expenses (in excess of Social Security). We assume that the contracts stipulate that no surrender value is paid out in case of policy cancellation. The (unpredictable) increase of medical costs in the future generates a systematic risk for the health insurance provider. Therefore, medical inflation is usually not guaranteed when setting the level premiums of the contracts at policy issue. Instead, premiums and eventually also reserves are regularly updated, accounting for observed medical inflation over the previous years.

In this chapter, we propose a simple but actuarially sound method that takes into account the observed medical inflation ex-post via a yearly recalculation of the premium levels. The premium-updating mechanism is based on a medical inflation index. This index quantifies the global increase in prices of medical goods and services and may thus differ from the classical consumer price index. Notice that a one-step version of the formula used in the present paper has been derived by Schneider (2002, Section 8) in the particular case of no reserve update. Here, this formula is extended to a multi-period setting, allowing for premium and/or reserve revisions.

Lifelong health insurance contracts and related premium updating mechanisms have been investigated in Vercruysse, Dhaene, Denuit, Pitacco and Antonio (2013) as well as

in Dhaene, Godecharle, Antonio and Denuit (2015). In the current chapter, we aim to derive a practical indexing method. Specifically, we show that indexing can be achieved by considering only premiums, without explicit reference to reserves. This appears to be relevant in practice as reserving mechanisms may not be transparent to policyholders and as some insurers do not compute contract-specific reserves, managing the whole portfolio in a collective way. The present study originates from a proposal for indexing medical insurance premiums in Belgium. As an application, we study the impact of various indexing rules on a typical medical insurance portfolio on the Belgian market.

This chapter is organized as follows. In Section 2.2, we describe the actuarial model for the health insurance contracts considered in this paper. Section 2.3 presents the one-step revision of the premium amount and/or of the accumulated reserve, as a consequence of medical inflation. Section 2.4 extends this one-step formula to periodic revisions during the coverage period. In Section 2.5, we consider the indexing mechanism recently proposed in Belgium. We show that the simple rule implemented after a Royal Decree dated March 2016 allows insurers to update premium amounts accounting for necessary reserve revaluations. The final sections 2.6 and 2.7 conclude the chapter.

2.2 Actuarial model

2.2.1 Two-decrement model

The origin of time is chosen at policy issue. Time t stands for the seniority of the policy (i.e., the time elapsed since policy issue). The policyholder's (integer) age at policy issue is denoted by x, so that upon survival at time k, he or she has reached age $x+k$. We denote the ultimate integer age by ω, assumed to be finite. This means that survival until integer age ω has a positive probability, whereas survival until integer age $\omega+1$ has probability zero.

We describe the lifelong health insurance policy considered in the previous section in a two-decrement Markov model, with states "active" (i.e. policy in force), "withdrawn" (i.e. policy has been cancelled) and "dead" (while the policy is in force), abbreviated as "a", "w" and "d", respectively. Let X_k be the status of the contract at time k, starting from $X_0 = \mathrm{a}$. The stochastic process $\{X_k,\ \ k = 0,1,2,\ldots\}$ describes the history of the contract.

For j and $k \in \{0,1,2,\ldots\}$, we define the sojourn (or non-exit) probability ${}_jp^{\mathrm{aa}}_{x+k}$ as

$$ {}_jp^{\mathrm{aa}}_{x+k} = P[X_{k+j} = \mathrm{a} | X_k = \mathrm{a}]. \tag{2.2.1}$$

In words, the quantity defined in (2.2.1) is the probability that a policy in force at age $x+k$ is still in force j years later. In accordance with standard actuarial notation, we omit the index j when it is equal to unity. The ultimate integer age ω is such that $p^{\mathrm{aa}}_{\omega-1} > 0$, while $p^{\mathrm{aa}}_{\omega} = 0$.

The probability that a policy in force at age $x+k$ has ceased j years later (due to death or withdrawal), is denoted by ${}_jp^{\mathrm{a}\bullet}_{x+k}$. This "exit" probability can be expressed as

$$ {}_jp^{\mathrm{a}\bullet}_{x+k} = P[X_{k+j} \neq \mathrm{a} | X_k = \mathrm{a}] = 1 - {}_jp^{\mathrm{aa}}_{x+k}. \tag{2.2.2}$$

We also introduce the probabilities ${}_jp^{\mathrm{ad}}_{x+k}$ and ${}_jp^{\mathrm{aw}}_{x+k}$, which are defined by

$$ {}_jp^{\mathrm{ad}}_{x+k} = P[X_{k+j} = d | X_k = \mathrm{a}] \text{ and } {}_jp^{\mathrm{aw}}_{x+k} = P[X_{k+j} = \mathrm{w} | X_k = \mathrm{a}]. \tag{2.2.3} $$

These are the probabilities of leaving the portfolio due to death and withdrawal, respectively, between ages $x+k$ and $x+k+j$.

2.2.2 Benefits and level premiums

The expected annual health cost at age $x+j$ which is denoted by b_{x+j}, is clearly random due to medical inflation. Let $b^{(0)}_{x+j}$ be an estimate at time 0 for the expected medical expenses in year $(0,1)$ for a person aged $x+j$ at time 0. At time 0, the insurer needs an estimate of the future costs for premium calculation, starting from the current expected cost $b^{(0)}_{x+j}$ for individuals aged $x+j$, increased by the assumed inflation rate. The insurer assumes a constant yearly medical inflation $f \geq 0$ over the coming years. Hence, $b^{(0)}_{x+j}$ $(1+f)^j$ is an estimate at time 0 of the expected medical expenses in year $(j, j+1)$ for a person aged $x+j$ in the beginning of that year. The results that we present hereafter can easily be generalized to the case of non-constant but deterministic estimates for future inflation in the coming years.

Throughout the chapter, a superscript "(k)", $k = 0, 1, 2, \ldots$, indicates that the quantity under consideration is based on information about medical costs available at time k. Hereafter, $\pi^{(k)}_{x,j}$ denotes the premium to be paid at time k for a contract that was underwritten at time $j \leq k$ at age x.

The tariff $\pi^{(0)}_{x,0}$ is determined from a technical basis, i.e. from assumptions about mortality, surrender, interest and medical inflation. In this chapter, we assume that, except for future medical inflation, the technical basis is guaranteed by the insurer. This means that the assumptions about mortality, surrender and interest rates are not subject to revision, and the corresponding risk is taken by the insurer. On the other hand, the uncertainty about future medical inflation levels induces systematic risk. Therefore, the amount of premium is revised ex-post on a yearly basis, using the observed inflation in the past year. We assume that in later years there is no update of the constant inflation scenario f that was used for premium calculation at time 0 so that the whole process is based on the tariff $\pi^{(0)}_{x,0}$ known at policy issue.

The level yearly premium $\pi^{(0)}_{x,0}$ for a health insurance contract underwritten at current time 0 on an insured aged x is determined by means of the equivalence principle. Let $v(0,j)$ be the discounting factor over the period $(0,j)$. The expected present value (or actuarial value) $B^{(0)}_x$ of the benefits paid by the insurer is then given by

$$ B^{(0)}_x = \sum_{j=0}^{\omega - x} b^{(0)}_{x+j} \; (1+f)^j \; {}_jE^{\mathrm{aa}}_x \tag{2.2.4} $$

where ${}_jE^{\mathrm{aa}}_x$ is the actuarial discounting factor accounting for mortality, lapses and interest,

over the period $(0, j)$, i.e.

$$ {}_jE_x^{\text{aa}} = v(0,j)\,{}_jp_x^{\text{aa}}. $$

Furthermore, let $\ddot{a}_x^{\text{aa}}$ be the actuarial value of an annuity-due paying a unit amount per year, as long as the policy is in force, i.e.

$$ \ddot{a}_x^{\text{aa}} = \sum_{j=0}^{\omega-x} {}_jE_x^{\text{aa}}. \tag{2.2.5} $$

We then have the level yearly premium

$$ \pi_{x,0}^{(0)} = \frac{B_x^{(0)}}{\ddot{a}_x^{\text{aa}}}. \tag{2.2.6} $$

In the following sections, we present an actuarially sound methodology for revising the level of the premium as inflation emerges over time. In order to ease the notations, we drop the superscript 'aa' for the actuarial discounting factor and the annuity value and simply denote ${}_jE_x^{\text{aa}}$ as ${}_jE_x$ and $\ddot{a}_x^{\text{aa}}$ as $\ddot{a}_x$.

2.3 Adapting the premium and/or the reserve level at time 1

2.3.1 Accumulated reserve

Suppose that we have arrived at time 1 and that the policy that was underwritten at age x at time 0 is still in force. This means that at time 1 a positive prospective reserve

$$ V_{x+1}^{(0)} = (1+f)\,B_{x+1}^{(0)} - \pi_{x,0}^{(0)}\,\ddot{a}_{x+1}, \tag{2.3.1} $$

is required for the insured now aged $x+1$, where $B_{x+1}^{(0)}$ and $\ddot{a}_{x+1}$ are defined similarly to (2.2.4) and (2.2.5), respectively.

Taking into account that the premium $\pi_{x,0}^{(0)}$ was determined via the equivalence principle (2.2.6), the prospective expression (2.3.1) for $V_{x+1}^{(0)}$ at time 1 can be transformed into the retrospective expression

$$ V_{x+1}^{(0)} = \left(\pi_{x,0}^{(0)} - b_x^{(0)}\right)({}_1E_x)^{-1} \tag{2.3.2} $$

which stands for the available reserve of the policyholder.

2.3.2 Revision of benefits

Suppose that the inflation for medical expenses observed during the first year is given by $f^{(1)}$. This means that at time 1, due to the observed medical inflation in the past year, the

expected annual medical expenses $b^{(0)}_{x+1+j}$ have to be updated to

$$b^{(1)}_{x+1+j} = \left(1+f^{(1)}\right) b^{(0)}_{x+1+j}, \qquad j = 0,1,2,\ldots \tag{2.3.3}$$

Notice that we assume in (2.3.3) that medical inflation over the past year is age-independent.

Taking into account this assumed uniformity of medical inflation over all ages, we find that at time 1, the actuarial value of future benefits $(1+f)B^{(0)}_{x+1}$, which was based on estimates available at time 0, has to be updated to

$$B^{(1)}_{x+1} = (1+f^{(1)})B^{(0)}_{x+1}. \tag{2.3.4}$$

2.3.3 Premium and/or reserve update

The required (prospective) reserve thus becomes

$$B^{(1)}_{x+1} - \pi^{(0)}_{x,0}\ \ddot{a}_{x+1},$$

which coincides with the available (retrospective) provision $V^{(0)}_{x+1}$ in (2.3.1) only if the observed inflation $f^{(1)}$ in the first year is equal to the assumed inflation f at time 0. This means that, due to the update of the actuarial value at age $x+1$ of future medical expenses, the available provision $V^{(0)}_{x+1}$ that was determined without knowing the observed medical inflation in the first year, turns out to be insufficient to cover future liabilities in case $f^{(1)} > f$. In order to restore the actuarial equivalence, the premium $\pi^{(0)}_{x,0}$ and/or the available provision $V^{(0)}_{x+1}$ will have to be updated to levels $\pi^{(1)}_{x,0}$ and $V^{(1)}_{x+1}$, respectively. Any pair $\left(V^{(1)}_{x+1},\ \pi^{(1)}_{x,0}\right)$ satisfying the equality

$$V^{(1)}_{x+1} = B^{(1)}_{x+1} - \pi^{(1)}_{x,0}\ \ddot{a}_{x+1} \tag{2.3.5}$$

will perform the task of resetting the actuarial equivalence. Hence, updating the premium and the available provision at time 1 can be performed in an infinite number of ways. Notice that (2.3.5) is the prospective reserve at time 1, based on updated benefits and premiums.

Subtracting (2.3.5) from (2.3.1), we find that for any pair $\left(V^{(1)}_{x+1},\ \pi^{(1)}_{x,0}\right)$ which restores the actuarial equivalence, the new premium level $\pi^{(1)}_{x,0}$ at time 1 is given by

$$\pi^{(1)}_{x,0} = \pi^{(0)}_{x,0} + \left(f^{(1)} - f\right)\ \pi^{(0)}_{x+1,0} - \frac{V^{(1)}_{x+1} - V^{(0)}_{x+1}}{\ddot{a}_{x+1}} \tag{2.3.6}$$

where

$$\pi^{(0)}_{x+1,0} = \frac{B^{(0)}_{x+1}}{\ddot{a}_{x+1}} \tag{2.3.7}$$

is the level premium at time 0 for a health insurance contract underwritten at that time for a person aged $x+1$.

Remark 2.3.1. *In the special case where $f = 0$ and the insurer decides to update the premium according to the observed medical inflation $f^{(1)}$, i.e.*

$$\pi_{x,0}^{(1)} = \left(1 + f^{(1)}\right) \pi_{x,0}^{(0)},$$

we find from (2.3.1), (2.3.4) and (2.3.5) that

$$V_{x+1}^{(1)} = \left(1 + f^{(1)}\right) V_{x+1}^{(0)}.$$

This means that in case no inflation is taken into account to determine the initial premium level $\pi_{x,0}^{(0)}$, indexing the premium according to the observed medical inflation $f^{(1)}$ requires the same proportional update of the available reserve.

2.3.4 Adapting the premium, only

Let us now assume that the level of the available provision is left unchanged, i.e.

$$V_{x+1}^{(0)} = V_{x+1}^{(1)}. \tag{2.3.8}$$

This means that the deviation of observed inflation $f^{(1)}$ from assumed inflation f in the first year is completely financed by the policyholder. From (2.3.6) it follows then that the new premium level at time 1 is given by

$$\pi_{x,0}^{(1)} = \pi_{x,0}^{(0)} + \left(f^{(1)} - f\right) \pi_{x+1,0}^{(0)}. \tag{2.3.9}$$

A similar formula has been obtained by Schneider (2002) in the particular case $f = 0$. Formula (2.3.9) shows that the premium increase $\pi_{x,0}^{(1)} - \pi_{x,0}^{(0)}$ at time 1 can be interpreted as the level premium corresponding to a "new" insurance contract underwritten at time 1 offering benefits with actuarial value equal to the benefit increases $\left(f^{(1)} - f\right) B_{x+1}^{(0)}$. This can be intuitively explained as follows: due to the increase in future medical costs from $(1+f)B_{x+1}^{(0)}$ to $(1+f^{(1)})B_{x+1}^{(0)}$, the policyholder now aged $x+1$ must virtually buy at time 1 a supplementary insurance policy, covering the benefit increase $\left(f^{(1)} - f\right) B_{x+1}^{(0)}$, whose price is $\left(f^{(1)} - f\right) \pi_{x+1,0}^{(0)}$ adding to $\pi_{x,0}^{(0)}$ in (2.3.9).

Formula (2.3.9) is a simple rule for updating the premium level at time 1: the new premium level $\pi_{x,0}^{(1)}$ follows from the original premium, the observed inflation over the past year and the insurer's tariff at time 0. The premium formula (2.3.9) can be rewritten in the following form:

$$\pi_{x,0}^{(1)} = \left(1 + \frac{\pi_{x+1,0}^{(0)}}{\pi_{x,0}^{(0)}} \left(f^{(1)} - f\right)\right) \pi_{x,0}^{(0)}.$$

This expression shows that the actual indexing for the original premium $\pi_{x,0}^{(0)}$ is

$$\left(f^{(1)} - f\right) \frac{\pi_{x+1,0}^{(0)}}{\pi_{x,0}^{(0)}}.$$

In case no inflation assumption is made at policy issue, i.e. $f = 0$, the proportional increase of the premium will be different (and usually higher) than the observed medical inflation $f^{(1)}$ over the first year. Also notice that in case the inflation assumption in the first year was too conservative, i.e. $f^{(1)} < f$, the premium level may be reduced at time 1.

2.4 Adapting the premium level at time k

2.4.1 Accumulated reserve

Suppose that we have arrived at time $k = 2,3,\ldots$ and that the policy that was underwritten on the person aged x at time 0 is still in force. The observed medical inflation up to time $k-1$ has been taken into account by restoring the actuarial equivalence and updating the premium levels at times $1,2,\ldots,k-1$. Suppose that the deviations of observed inflation from assumed inflation f are completely financed by the policyholder, which means that the available provisions are not updated. Let $V_{x+k-1}^{(k-1)}$ and $\pi_{x,0}^{(k-1)}$ be the available provision and the premium level determined at time $k-1$. They were set such that the actuarial equivalence at time $k-1$ was restored:

$$V_{x+k-1}^{(k-1)} = B_{x+k-1}^{(k-1)} - \pi_{x,0}^{(k-1)} \ddot{a}_{x+k-1}. \tag{2.4.1}$$

In this formula, $B_{x+k-1}^{(k-1)}$ is the actuarial value at time $k-1$ of the future health benefits related to an insured of age $x+k-1$ at that time, i.e.

$$B_{x+k-1}^{(k-1)} = \sum_{j=0}^{\omega-x-k+1} b_{x+k-1+j}^{(k-1)} \; (1+f)^j \; {}_jE_{x+k-1},$$

where $b_{x+k-1+j}^{(k-1)}$ is the expected health benefit in year $(k-1,k)$ for a person aged $x+k-1+j$ in the beginning of that year, based on the information available at time $k-1$:

$$b_{x+k-1+j}^{(k-1)} = b_{x+k-1+j}^{(0)} \prod_{l=1}^{k-1} (1+f^{(l)}).$$

Notice that we assume that inflation is age-independent. Furthermore, the ${}_jE_{x+k-1}$ are the appropriate actuarial discount factors, accounting for mortality, lapses and interest, while $\ddot{a}_{x+k-1}$ is an annuity-due paying an amount of 1 per year to the insured with current age $x+k-1$, as long as the policy remains in force.

The available provision at time k for the policy still in force at that time is then given by

$$V_{x+k}^{(k-1)} = \left(V_{x+k-1}^{(k-1)} + \pi_{x,0}^{(k-1)} - b_{x+k-1}^{(k-1)}\right) ({}_1E_{x+k-1})^{-1}. \tag{2.4.2}$$

The available provision acts as a savings account, which first builds up by the premium surpluses in the early years (when $\pi_{x,0}^{(k-1)} > b_{x+k-1}^{(k-1)}$), whereas it melts away in later years due to the premium shortfalls in these years (when $\pi_{x,0}^{(k-1)} < b_{x+k-1}^{(k-1)}$).

Taking into account the restored actuarial equivalence (2.4.1) at time $k-1$, the available reserve $V_{x+k}^{(k-1)}$ at time k can be expressed in the following prospective form:

$$V_{x+k}^{(k-1)} = (1+f)\, B_{x+k}^{(k-1)} - \pi_{x,0}^{(k-1)}\ \ddot{a}_{x+k}, \tag{2.4.3}$$

with

$$B_{x+k}^{(k-1)} = \sum_{j=0}^{\omega-x-k} b_{x+k+j}^{(k-1)}\ (1+f)^j\ {}_jE_{x+k}.$$

2.4.2 Premium update

Due to the observed medical inflation during the k-th year, the actuarial value of future health benefits $(1+f)\, B_{x+k}^{(k-1)}$ based on an evaluation at time $k-1$ has to be updated to $B_{x+k}^{(k)}$ which, under the age-uniform medical inflation $f^{(k)} \geq 0$, is given by

$$B_{x+k}^{(k)} = (1+f^{(k)}) B_{x+k}^{(k-1)}.$$

At time k, the premium level $\pi_{x,0}^{(k-1)}$ and/or the available provision $V_{x+k}^{(k-1)}$ have to be replaced by $\pi_{x,0}^{(k)}$ and $V_{x+k}^{(k)}$, respectively, in order to restore the actuarial equivalence:

$$V_{x+k}^{(k)} = B_{x+k}^{(k)} - \pi_{x,0}^{(k)}\ \ddot{a}_{x+k}. \tag{2.4.4}$$

From (2.4.3) and (2.4.4), we find that for any actuarial equivalence restoring pair $\left(V_{x+k}^{(k)}, \pi_{x,0}^{(k)}\right)$, the updated premium $\pi_{x,0}^{(k)}$ can be expressed as

$$\pi_{x,0}^{(k)} = \pi_{x,0}^{(k-1)} + \left(f^{(k)} - f\right)\ \pi_{x+k,k-1}^{(k-1)} - \frac{V_{x+k}^{(k)} - V_{x+k}^{(k-1)}}{\ddot{a}_{x+k}}, \tag{2.4.5}$$

where $\pi_{x+k,k-1}^{(k-1)}$ is given by

$$\pi_{x+k,k-1}^{(k-1)} = \frac{B_{x+k}^{(k-1)}}{\ddot{a}_{x+k}}, \tag{2.4.6}$$

which is the initial level premium for a lifelong health insurance contract underwritten at time $k-1$ on a person of age $x+k$ at that time.

Assuming again that the observed inflation $f^{(k)}$ is solely financed by the policyholder, i.e.

$$V_{x+k}^{(k)} = V_{x+k}^{(k-1)},$$

the premium updating formula (2.4.5) reduces to

$$\pi_{x,0}^{(k)} = \pi_{x,0}^{(k-1)} + \left(f^{(k)} - f\right) \pi_{x+k,k-1}^{(k-1)}. \tag{2.4.7}$$

Hence, the updated premium $\pi_{x,0}^{(k)}$ at time k is equal to the premium $\pi_{x,0}^{(k-1)}$ paid the year before, augmented by the product of the medical inflation deviation $\left(f^{(k)} - f\right)$ observed over the past year and last year's premium for a new contract that was issued on a person of age $x+k$.

The updated premium $\pi_{x,0}^{(k)}$ can also be written as

$$\pi_{x,0}^{(k)} = \left(1 + \frac{\pi_{x+k,k-1}^{(k-1)}}{\pi_{x,0}^{(k-1)}} \left(f^{(k)} - f\right)\right) \pi_{x,0}^{(k-1)}. \tag{2.4.8}$$

This expression clearly shows that the proportional premium increase at time k is different from the medical inflation deviation $\left(f^{(k)} - f\right)$ that was revealed over the past year. Obviously, the proportional premium increase depends on the age x at policy issue as well as on the number k of years that the contract has been in force so far. The proportional increase of the premium will usually be larger for policies that are longer in force.

From (2.4.7) which holds for $k = 1,2,3,\ldots$, we find that

$$\pi_{x,0}^{(k)} = \pi_{x,0}^{(0)} + \sum_{j=1}^{k} \left(f^{(j)} - f\right) \pi_{x+j,j-1}^{(j-1)}, \tag{2.4.9}$$

with

$$\pi_{x+j,j-1}^{(j-1)} = \frac{B_{x+j}^{(j-1)}}{\ddot{a}_{x+j}}. \tag{2.4.10}$$

Formula (2.4.9) has an intuitive interpretation. Indeed, the premium level $\pi_{x,0}^{(k)}$ to be paid at time k is equal to the initial premium level $\pi_{x,0}^{(0)}$, augmented with the extra premia for all the virtually added contracts covering the increases in medical costs in any of the first k years.

Using the assumption of an age-uniform medical inflation in each of the past years, we find that

$$B_{x+j}^{(j-1)} = B_{x+j}^{(0)} \prod_{l=1}^{j-1} \left(1 + f^{(l)}\right), \qquad j = 1,2,3,\ldots, \tag{2.4.11}$$

provided we set $\prod_{l=1}^{0}(1+f^{(l)})=1$, by convention. Taking into account (2.4.10), the expression above immediately leads to

$$\pi_{x+j,j-1}^{(j-1)} = \pi_{x+j,0}^{(0)} \prod_{l=1}^{j-1} \left(1+f^{(l)}\right), \qquad j=1,2,3,\ldots \tag{2.4.12}$$

It follows then from (2.4.9) that the updated premium level $\pi_{x,0}^{(k)}$ at time k can be written as

$$\pi_{x,0}^{(k)} = \pi_{x,0}^{(0)} + \sum_{j=1}^{k} \left(f^{(j)}-f\right) \pi_{x+j,0}^{(0)} \prod_{l=1}^{j-1}(1+f^{(l)}). \tag{2.4.13}$$

This is an expression for the updated premium level $\pi_{x,0}^{(k)}$ at time k for a contract underwritten to a person aged x at time 0, in terms of the observed inflation levels $f^{(1)}, f^{(2)}, ..., f^{(k)}$ in the past years and the insurer's tariff $\pi_{y,0}^{(0)}$ at policy issue.

2.4.3 Adaptation to age-specific inflation

In this chapter we made the simplifying assumption that in any year $k=1,2,\ldots,$ observed inflation is uniform over all ages, i.e.

$$b_{x+j}^{(k)} = (1+f^{(k)})\, b_{x+j}^{(k-1)}, \qquad j=1,2,\ldots,$$

for some age-independent inflation factor $f^{(k)}$. Notice however that the results presented here can in a straightforward way be adapted to the case of age-specific medical inflation (see Denuit et al. 2017) by replacing the inflation factor $f^{(k)}$ in the formula above by an age-dependent factor. In this case, we have that

$$b_{x+j}^{(k)} = (1+f_{x+j}^{(k)})\, b_{x+j}^{(k-1)} \qquad j=1,2,\ldots$$

Once the age-specific inflation factors $f_{x+j}^{(k)}$ have been set, we can determine the global inflation factors $\overline{f}_{x+k}^{(k)}$ from

$$B_{x+k}^{(k)} = (1+\overline{f}_{x+k}^{(k)})\, B_{x+k}^{(k-1)}.$$

Remark that the interpretation of the factors $\overline{f}_{x+k}^{(k)}$ is not straightforward, as it is a weighted average of the observed inflation factors for all ages from $x+k$, with weights that depend on age-specific expected health benefits and actuarial discount factors.

In the generalized setting with age-dependent inflation, the simple premium updating obtained in this chapter has to be replaced by

$$\pi_{x,0}^{(k)} = \pi_{x,0}^{(k-1)} + \left(\overline{f}_{x+k}^{(k)}-f\right)\ \pi_{x+k,k-1}^{(k-1)},$$

with a similar interpretation as before: the updated premium $\pi_{x,0}^{(k)}$ at time k is equal to the premium $\pi_{x,0}^{(k-1)}$ paid the year before, augmented by the product of the deviation of global medical inflation factor $\overline{f}_{x+k}^{(k)}$ from the assumed inflation f and the initial premium for a new contract that was issued the year before on a person of age $x+k$.

2.5 Case study: The new indexing mechanism for the Belgian medical insurance market

2.5.1 Indexing rule imposed by the Belgian law

Individual private coverages are lifelong by law. In case of level premiums, the initial premium amount is fixed at policy issue and then linked to the CPI or to a specific medical index. The Federal Agency KCE studied different indexing mechanisms, see Devolder et al. (2008). The Royal Decree defining the premium indexing mechanism to be applied by insurance companies operating in Belgium has been cancelled on December 29, 2011, one of the reasons being that the updating mechanism for the premiums to adjust for observed but unanticipated inflation did not take into account the shortfall of the accumulated reserves.

Recently, a Belgian Royal Decree dated March 18, 2016 introduced a new updating mechanism for individual private coverages. The newly proposed mechanism, which is intended to be both appropriate for the insurers and transparent towards the clients, is given by

$$\pi_{x,0}^{(k)} = \left(1 + 1.5\ f^{(k)}\right) \pi_{x,0}^{(k-1)}, \tag{2.5.1}$$

subject to some restrictions that are not be considered in the present paper (as they only apply to very special cases, not encountered in our numerical examples). Here it is assumed that the level premiums are determined without assuming future inflation: in all our numerical illustrations, we always take $f = 0$. Henceforth, the premiums calculated according to (2.5.1) are called the "1.5 rule premiums" and denoted by $\pi_{x,0}^{(k)}\,(150\%)$. We compare these premiums with the "exact premiums", which follow from (2.4.8):

$$\pi_{x,0}^{(k)} = \left(1 + \alpha_{x,0}^{(k)}\ f^{(k)}\right) \pi_{x,0}^{(k-1)}, \tag{2.5.2}$$

with

$$\alpha_{x,0}^{(k)} = \frac{\pi_{x+k,k-1}^{(k-1)}}{\pi_{x,0}^{(k-1)}}, \tag{2.5.3}$$

which holds under the assumption that $f = 0$ and that reserves are not updated, i.e. $V_{x+k}^{(k-1)} = V_{x+k}^{(k)}$. In order to distinguish the exact premiums (2.5.2) from the premiums derived from the 1.5 rule, we often denote them by $\pi_{x,0}^{(k)}\,(\text{exact})$.

Comparing (2.5.1) and (2.5.2), we see that the new Belgian regulation amounts to replacing the "exact" updating factors $\alpha_{x,0}^{(k)}$ by a single number, namely 1.5. Hereafter, we assess the impact of using the simplified mechanism (2.5.1) instead of the exact mechanism (2.5.2). In particular, we compare the indexing factors $\alpha_x^{(k)}$ with the constant 1.5 in order to investigate whether using 1.5 instead of the correct indexing factor appears to be a sufficiently prudent approach for the insurance company and at the same time, a not too conservative approach for the insured. Let us mention that (2.5.1) determines the maximum premium update allowed by the law, so that we mainly adopt the insurer's point of view and examine whether a rule like (2.5.1) may threaten its solvency or not.

2.5.2 Technical basis

The contract studied is a unisex lifetime hospitalization cover, i.e. men and women combined. The ultimate age is set at 110. The minimum subscription age is set at 25 years. The numerical examples do not consider the possibility of lapse, with reference to the current situation on the Belgian market. The expected claims costs are shown in Figure (2.1). They are based on the annual claims amounts observed according to age, with an extrapolation to advanced ages. We can recognize the bump associated with accidents and childbirth between 20 and 40 years. Costs have been normalized to reflect the average annual cost of hospitalization provided by Mutualité Chrétienne. The assumed discount factors correspond to a constant yearly interest rate $i = 1\%$. In terms of mortality, prospective mortality tables of the Bureau Fédéral du Plan are used. We assume that experienced interest rates, mortality rates and withdrawal rates are equal to their corresponding values in the technical basis. Furthermore, we assume an observed medical inflation of 2 percent, i.e. $f^{(k)} = 2\%$ for all $k = 1, 2, \ldots$

2.5.3 Effect of inflation

We consider a policyholder underwriting a hospitalization insurance policy on January 1, 2017, at the age of 25. We assume constant medical inflation of 2% per year to simplify our presentation. In the following, we will also examine the effect of age at underwriting, considering older policyholders, and the rate of inflation. Table (2.1) describes the details of the calculation of the exact increase in the premium resulting from inflation of 2% per year, based on the above assumptions. For each year, the following are included:

- the level premium, payable annually in advance, corresponding to a contract starting at the attained age, taken out during the previous year;
- the new premium broken down into the sum of the previous premium plus inflation multiplied by the premium at the age reached (therefore the cost of the virtual contract covering the increase due to inflation);
- the rate of increase of the premium and

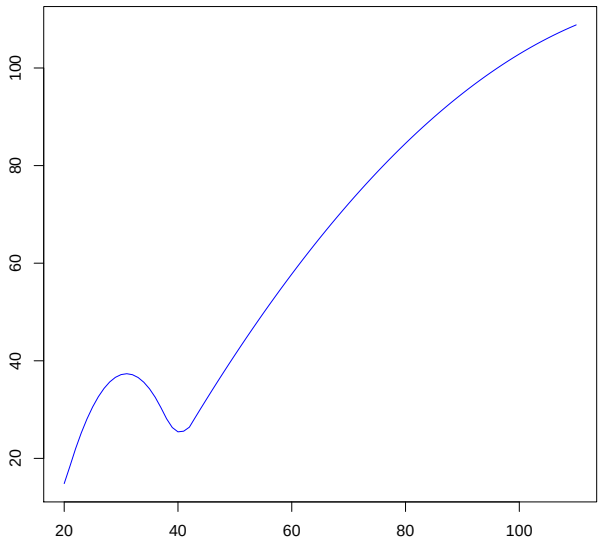

Figure 2.1: Average annual amount (euro) assumed for the insurer's services, depending on the age reached.

Year	contract seniority	(a)	(b)	(c)	(d)	Increase rate	Premium 150%
2017	0	-	-	-	49.802	-	-
2018	1	51.2	49.8	1.02	50.8	2.06%	51.3
2019	2	52.6	50.8	1.05	51.9	2.07%	52.8
2020	3	54.1	51.9	1.08	53.0	2.08%	54.4
2021	4	55.6	53.0	1.11	54.1	2.10%	56.1
2022	5	57.1	54.1	1.14	55.2	2.11%	57.7
2023	6	58.6	55.2	1.17	56.4	2.12%	59.5
2024	7	60.2	56.4	1.20	57.6	2.13%	61.3
2025	8	61.8	57.6	1.24	58.8	2.15%	63.1
2030	13	71.6	64.1	1.43	65.6	2.23%	73.1
2035	18	85.1	71.8	1.7	73.5	2.37%	84.8
2040	23	101.0	80.9	2.02	83.0	2.50%	98.3

Table 2.1: Detailed calculation of the increase in the annual premium over time. (a)=premium at reached age, (b)=previous premium, (c)= inflation $\times$ (a), (d)=adapted premium=(b)+(c)

- the maximum premium induced by the 150% rule (under the assumption that the insurer passes on each year 1.5 times the inflation observed on the previous value in the same column).

Thus, in the example illustrated in Table 2.1, the level subscription premium is 49,802 and does not take into account any possible inflation of future benefits. In the first year, the observed medical inflation is 2% and the premium of 49,802 must therefore be increased to compensate for the increase in expected medical benefits. The new adequate premium will not be a simple increase of the previous premium of 2%, which would be insufficient. An additional premium will be added to the previous premium and it will cover the increase in expected benefits. The amount of this additional premium is that of a "virtual" contract starting the following year and covering the part of the costs not initially foreseen. The adjusted premium (b + c) can therefore be obtained as the sum between the previous premium and the inflation observed for the past year multiplied by the premium corresponding to the age reached.

The growth rate shown in the table corresponds to the increase in the previous premium, and we see that it exceeds the inflation rate for the past year (i.e. 2% in the example). The correction factor to be applied to the inflation rate for the past year, which we will henceforth call the indexing factor, is equal to the ratio between the premium of a contract initiated at the age reached and the premium for the previous. This factor is usually greater than one since the premium for a contract starting one year later is usually higher than the base premium. This therefore shows that the new premium will be higher

than the previous premium corrected on the basis of the inflation of the past year, inducing a growth rate higher than the inflation rate of the past year.

The last column of Table 2.1 illustrates the maximum premium authorized by the Royal Decree of March 2016, i.e. the previous premium (in the same column) increased by 1.5 times the observed inflation (which amounts to 2% in our example) . This column therefore corresponds to the evolution of the premium with an insurer who would apply each year the maximum adjustment authorized by the legislator. This maximum premium is always much higher, in our case, than the technically adequate premium taking into account the indexation.

We note that the insurer should in principle not pass on the 150% of the inflation observed as the royal decree authorizes. Indirectly, the indexation mechanism could therefore help to compensate for other technical losses linked to the discounting, mortality or termination of contracts, for example, by avoiding sudden corrections of the premium in the event of deviation from other components of the technical bases, endangering the solvency of the insurer.

The exact indexing factors corresponding to different scenarios are shown in Figures 2.2, 2.3 and 2.4, in order to make a comparison with the 150% rule and challenge that it is prudent. The abscissa of the graphs corresponds to the length of the contract. The scenarios considered apply variable initial ages (40 and 65 years at subscription, rather than 25 years as previously, in Figure 2.2), variable observed inflation rates (passing to 1% or 3% instead of 2 %, in Figure 2.3) and variable interest rates (ranging from 0 to 2 % in Figure 2.4). The aim is to detect situations where the 150% rule is insufficient.

In Figure 2.2 we see that the indexing factor for a 65-year-old individual is the lowest. The ratio between the premium at attained age and the previous premium is therefore lower: it may even become less than unity at very advanced ages. At such ages, in fact, the high probabilities of death outweigh the increased costs expected. There is no excess over level 1.5 at all the ages considered.

In Figure 2.3, we see that from a sufficient length of contract, the situation which induces the highest corrective factor is an inflation of 1%. Since the corrective factor is the ratio between the premium at the age reached and the premium for the previous year, low inflation reduces the numerator to a lesser extent than it reduces the denominator, for older ages. The 150% rule seems prudent enough given that the indexation factor remains below 1.5.

In Figure 2.4, we depict the behavior of the exact indexing factor for 3 different constant interest rate scenarios considered: 0-1-2%. In our numerical analysis the 1.5 rule again stands out as being overall conservative as the indexation factor remains always below 1.5. We see that the situation which induces the highest corrective factor is an interest rate of 2%. We refer the reader to the similar results displayed in Figure 6 of Denuit et al. (2017).

In Table 2.1 a base case scenario of constant 2% inflation is proposed. In addition to the constant inflation rate scenario, the analysis is performed in the advent of a permanent (see Figure 2.5 and Table 2.2) or temporary jump (see Figure 2.6 and Table 2.3). A time $t = 3$ inflation jumps at 5% and the jump is supposed to be permanent or temporary:

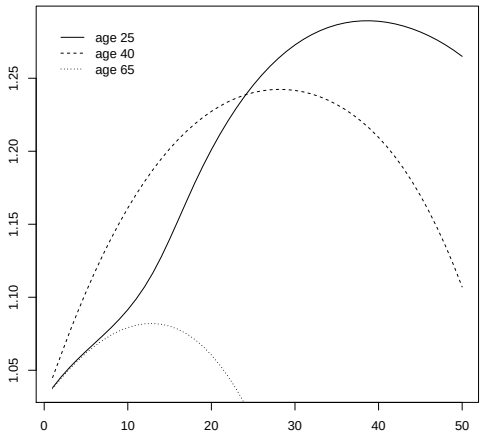

Figure 2.2: Evolution of the indexation factor according to the different subscription age scenarios.

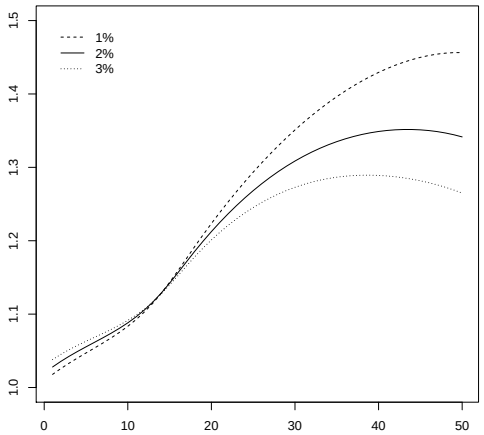

Figure 2.3: Evolution of the indexation factor according to the different scenarios of long-term medical inflation, for an insured aged 25 at the time of subscription.

Year	Inflation	(a)	(b)	(c)	(d)	Increase rate	Premium 150%
0	-	-	-	-	49.802	-	-
1	2%	51.2	49.8	1.02	50.8	2.06%	51.3
2	2%	52.6	50.8	1.05	51.9	2.07%	52.8
3	2%	54.1	51.9	1.08	53.0	2.08%	53.4
4	5%	57.2	53.0	2.86	55.8	5.40%	56.9
5	5%	60.4	55.8	3.02	58.8	5.42%	60.0
6	5%	63.9	58.8	3.20	62.0	5.43%	63.3
7	5%	67.6	62.0	3.38	65.4	5.45%	66.7
8	5%	71.5	65.4	3.57	69.0	5.46%	70.3
13	5%	95.7	85.5	4.78	90.3	5.59%	91.9
18	5%	131.9	112.7	6.59	119.4	5.85%	121.2
23	5%	180.4	150.3	9.02	159.3	6.00%	161.5

Table 2.2: Detailed calculation of the increase in the annual premium over time with a permanent inflation jump. (a)=premium at reached age, (b)=previous premium, (c)= inflation $\times$ (a), (d)=adapted premium=(b)+(c)

- Permanent jump : 2%,2%,2%,5%,5%,5%,...
- Temporary jump : 2%,2%,2%,5%,2%,2%,...

The prudency of the 1.5 rule is confirmed in all the different 'base case' sensitivity analyses. We should also point out a dynamic feature of the correcting factor (see Hanbali et al. 2017). At time 1, one year after the conclusion of this contract, the insurer must adjust the future premiums in order to integrate the information observed during that year, and must also take into account the previous premium paid. Since only one premium has been paid, the correcting factor will often be relatively low in the first year and the 1.5 approximation will overestimate its exact value. From this we can deduce that the 1.5 rule is prudent and overestimates the adaptation during the first few years of coverage during which the reserve is not too high. However, after some years the exact value of the indexing factor could exceed this approximation and lead - depending on the evolution of the parameters on which this factor depends - to an underestimation. However, it should be noted that that in the event of an underestimation of the adaptation, the insurer can always ask for an authorisation to change the premium from the NBB.

2.6 Future perspectives

2.6.1 Risk transfer mechanism

In this chapter, the proposed solution consists in fully transferring the medical risk from the insurer to policyholder. The medical inflation risk can also be transferred to the financial market through securitisation or to the working population via social security (see the 'Discussions and Extensions' Chapter for a discussion on that latter option).

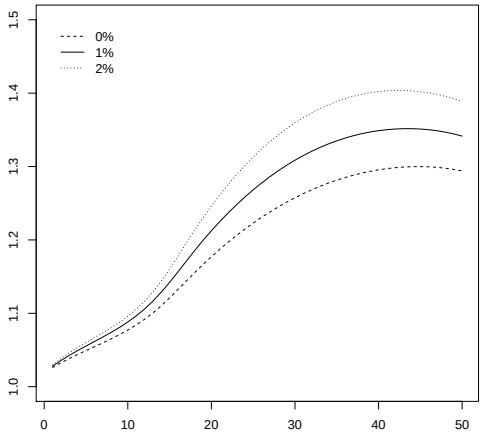

Figure 2.4: Evolution of the indexation factor according to the different scenarios of interest rate, for an insured aged 25 at the time of subscription.

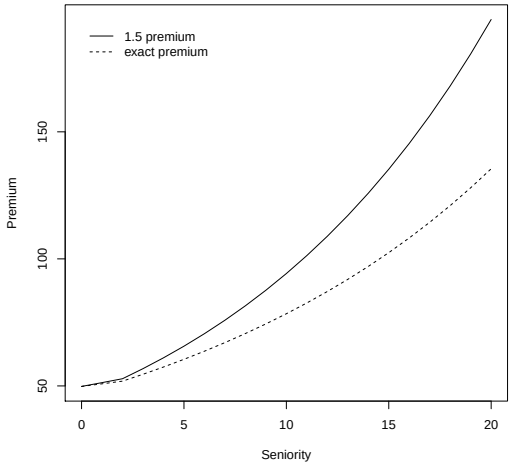

Figure 2.5: Evolution of the premium with a permanent inflation jump

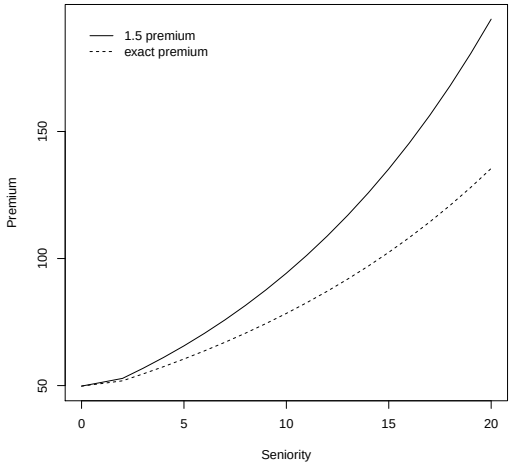

Figure 2.6: Evolution of the premium with a temporary inflation jump

Year	Inflation	(a)	(b)	(c)	(d)	Increase rate	Premium 150%
0	-	-	-	-	49.802	-	-
1	2%	50.8	49.8	1.02	50.8	2.04%	51.3
2	2%	51.9	50.8	1.04	51.9	2.04%	52.3
3	2%	54.6	51.9	1.09	52.9	2.10%	53.4
4	5%	55.7	52.9	2.79	55.7	5.26%	56.9
5	2%	56.9	55.7	1.14	56.9	2.04%	57.4
6	2%	58.1	56.9	1.16	58.0	2.04%	58.6
7	2%	59.3	58.0	1.19	59.2	2.04%	59.8
8	2%	60.6	59.2	1.21	60.4	2.05%	61.0
13	2%	67.6	65.5	1.35	66.9	2.06%	67.5
18	2%	75.7	72.6	1.51	74.1	2.09%	74.8
23	2%	85.5	80.6	1.71	82.3	2.12%	83.0

Table 2.3: Detailed calculation of the increase in the annual premium over time with a temporary inflation jump. (a)=premium at reached age, (b)=previous premium, (c)= inflation $\times$ (a), (d)=adapted premium=(b)+(c)

Medical inflation bonds or morbidity-linked securities are still currently hypothetical. Yet the use of securitisation could enable to hedge oneself against medical inflation or morbidity risk. Securitisation through capital markets, could play an important role, offering additional capacity and liquidity to the market (see Biffis et al. 2010).

For its part inflation-linked bonds do exist and were created several decades ago (see Kherkov 2005). There are indeed zero coupon inflation-indexed swaps or also inflation-linked bonds, which have CPI as an underlying, e.g. to index their coupons or notional. But medical inflation is typically higher than CPI.

There are also theoretical discussions about GDP-linked bonds where the coupon would depend on evolutions in economic growth. In this way, the national debt could be stabilised in, for example, COVID-19 crisis periods. Since the academic literature (e.g. Getzen 2020) and the Belgian 'Bureau Federal du Plan' are assuming that medical costs move in line with GDP, GDP-linked bonds could be a first proxy for medical inflation indexed bonds. There are already some papers to determine the pricing of such instruments, but these instruments are only theoretical and therefore they are not traded.

Even if there is no financial instrument linked specifically to the evolution of medical costs, there could exist over-the-counter basic inflation swaps for medical inflation that are customised at the request of an institution by an investment bank, but these are not traded on the stock exchange and are certainly not liquid.

2.6.2 Risk factors

Apart from the medical inflation risk, other risk factors could be considered, like interest rate. The aim in this chapter was to challenge the indexation mechanism defined in Belgian law, i.e. to challenge that the new premium only depends on the initial tariff and medical inflation.

If only the inflation risk is considered as in this chapter, the benefits $B_{x+k}^{(k-1)}$ are simply increased by the inflation factor $(1+f^{(k)})$. We can premium equivalence and obtain $\pi_{x,0}^{(k)}$ and $V_{x+k}^{(k)}$ such that the following holds:

$$\pi_{x,0}^{(k)} = \pi_{x,0}^{(k-1)} + \frac{B_{x+k}^{(k)} - B_{x+k}^{(k-1)}}{\overline{a}_{x+k}} - \frac{V_{x+k}^{(k)} - V_{x+k}^{(k-1)}}{\overline{a}_{x+k}} \tag{2.6.1}$$

In this chapter we have assumed that actual interest, lapse and mortality rates remain equal to their assumed values entering actuarial formulas. In practise these assumptions could be violated and they should also be revised periodically. Changes in the hypotheses, like interest rate, lapse or mortality, induce a modification of prospective reserves. The changes should be recuperated on premiums or on constituted reserves.

In order to maintain the basic equivalence principle, retrospective reserve should equal the prospective one, i.e.

$$V_{x+k}^{\text{retro}} = V_{x+k}^{\text{prosp}}$$

for all k. This relation can be used all the time and for any component of the technical basis. The nice simplifications obtained in this chapter, with no explicit reference to reserve but only premiums at initiation, are only possible with the medical inflation factor as proposed by the Belgian law.

If there is a change of interest rate from r^{old} to r^{new} at time k inducing a premium adaptation from π_{x+k}^{old} to π_{x+k}^{new}, we have :

$$V_{x+k}^{\text{retro}} = \sum_{l=1}^{\omega-x-k} {}_lp_{x+k}^{\text{aa}} \frac{b_{x+k+l}^{(k)} - \pi_{x+k}^{\text{old}}}{(1+r^{\text{old}})^k}$$

with $b_{x+k}^{(k)}$ the expected medical cost at age $x+k$ computed at time k and

$$V_x^{\text{prosp}} = \sum_{l=1}^{\omega-x-k} {}_lp_{x+k}^{\text{aa}} \frac{b_{x+k+l}^{(k)} - \pi_{x+k}^{\text{new}}}{(1+r^{\text{new}})^k}.$$

We obtain

$$\pi_{x+k}^{\text{new}} = \left(\sum_{l=1}^{\omega-x-k} {}_lp_{x+k}^{\text{aa}} \frac{b_{x+k+l}^{(k)}}{(1+r^{\text{new}})^k} - V_{x+k}^{\text{retro}} \right) \Big/ \left(\sum_{l=1}^{\omega-x-k} {}_lp_{x+k}^{\text{aa}} \frac{1}{(1+r^{\text{new}})^k} \right)$$

and the resulting update mechanism is not intuitive and includes an explicit reference to the available reserve V_{x+k}^{retro}.

2.7 Conclusion

As an accurate prediction of future medical inflation is practically impossible, an insurer selling lifelong health insurance coverage usually does not make a guaranteed assumption concerning future inflation at policy issue, in order to avoid the risk of underestimating this inflation. Moreover, the systematic nature of medical inflation, affecting each policy in the same direction, implies that the Law of Large Numbers, which is the crucial concept on which insurance business is built, is not applicable. As a consequence, in lifelong health insurance, the uncertainty concerning medical inflation usually remains with the insureds, who will pay variable future premiums which are directly related to the level of inflation that will emerge over time.

In this chapter, we described a relatively simple but actuarially adequate individual updating mechanism (2.4.7), which can also be expressed as (2.4.8), for such lifelong health insurance contracts. The premium level is yearly updated, taking into account the observed inflation over the past year. From formula (2.4.8) it follows that the required proportional increase of the premium does not only depend on the difference between observed and assumed medical inflation in the previous year, but also on the age at policy issue and on the time since policy issue.

The analysis carried out in this work show that the "150%" rule, constituting a maximum for the insurer, makes it possible to take into account the necessary revaluation of reserves. The actual increases in premiums should remain well below the maximum authorized, provided that the other assumptions (discount rate, mortality, etc.) do not generate technical losses. The latter may indirectly be covered by the insurer through the indexation mechanism provided for by the legislator. In reality, the above-mentioned assumptions will be chosen in a conservative way, implying that the insurer will very likely make technical gains. These technical gains might be (partly) redistributed to the insureds via an increase of the available provisions, implying a partial financing of the observed medical inflation by the insurer.

Although open to criticism from the point of view of strict actuarial technique, the new indexation mechanism has the great advantage of simplicity and transparency. It should make it possible to avoid sudden increases in premiums in the future, by making them evolve according to the observed inflation in the cost of hospitalizations, thus bringing serenity to a market which has experienced many upheavals in recent years. Finally, it should be noted that the insurer is in no way responsible for the increase in health care costs. He can only note it and pass it on to premiums, otherwise he will no longer be able to honor his commitments to policyholders.

References

- Biffis E. & Blake D. (2010) Mortality-linked securities and derivatives. In Optimizing the Aging, Retirement and Pensions Dilemma. John Wiley & Sons
- Biffis, E. & Blake, D. (2014) Keeping some skin in the game: How to start a capital market in longevity risk transfers. North American Actuarial Journal 18(1), 14-21
- Devolder, P., Denuit, M., Maréchal, X., Yerna, B.-L., Closon, J.-P., Léonard, C., Senn, A. & Vinck, I. (2008). Construction d'un index médical pour les contrats privés d'assurance maladie. KCE Reports - Centre Fédéral d'Expertise des Soins de Santé, Volume 96B. Available from http://www.centredexpertise.fgov.be
- Dhaene, J., Godecharle, E., Antonio, K., Denuit, M. & Hanbali, H. (2017). Lifelong Health Insurance Covers with Surrender Values: Updating Mechanisms in the Presence of Medical Inflation. Astin Bulletin, 47(3), 803-836.
- Dickson, D. C., Hardy, M. R. & Waters, H. R. (2013). Actuarial Mathematics for Life Contingent Risks. Cambridge University Press.
- Getzen, T. (2020) Measuring National Health Expenditures. Available at SSRN 3564784.
- Haberman, S. & Pitacco, E. (1999). Actuarial Models for Disability Insurance. CRC Press.
- Hanbali H., Claassens H., Denuit M., Dhaene J. & Trufin J. (2017) Application de l'indice médical dans les contrats d'assurance maladie en Belgique. Research report AFI-17114, KU Leuven.
- Kherkov, J. (2005) Inflation Derivatives Explained Markets, Products, and Pricing. Lehman Brothers.
- Pitacco, E. (1999). Multistate models for long-term care insurance and related indexing problems. Applied Stochastic Models in Business and Industry 15, 429-441.
- Schneider, E. (2002). The main features of German private health insurance. In the Proceedings of the 27th ICA Health Seminar, Cancun.
- Vercruysse, W., Dhaene, J., Denuit, M., Pitacco, E. & Antonio, K. (2013). Premium indexing in lifelong health insurance. Far East Journal of Mathematical Sciences 2013, 365-384.
- Wynand, P. M. M., De Ven, V. & Ellis, R. P. (2000). Risk adjustment in competitive health plan markets. In Handbook of health economics 1, 755-845. Elsevier.
- Zhao, B. (2012) A modified Lee-Carter model for analysing short-base-period data. Population Studies 66(1), 39-52.

Chapter 3

Stochastic modelization of claim costs trends

The present chapter is based on the following published paper:

- Christiansen, M., Denuit, M., Lucas, N., Schmidt, J. P. (2018). Projection models for health expenses. Annals of Actuarial Science, 12(1), 185-203.

3.1 Introduction

Morbidity rates and mortality rates often share a very similar age pattern, with higher values around birth and at young adult ages (near the so-called accident hump), and then monotonically increasing at older ages, first exponentially before switching to an increasing concave behavior. The same structure is found for the expected number of claims in sickness or medical insurance, with some peculiarities (such as the hump induced by childbearing for young women). Corresponding yearly insurance claim costs, being influenced by their frequency component, also exhibit a similar age shape. This suggests that models developed to describe the age structure of mortality can be useful for these related quantities, too.

Inspired by the classical log-bilinear mortality projection model proposed by Lee and Carter (1992), we show in this chapter how to adapt mortality studies to describe trends in medical expenses.

The Lee-Carter model has been successfully applied to project disability rates, in addition to mortality ones, in Christiansen et al. (2012) as well as in Levantesi and Menzietti (2012). For another approach based on parametric models, see Renshaw and Haberman (2000). The aim of this chapter is to address the main differences between mortality and medical expenses. In both cases, a stable age-specific pattern is found consistently over time and short-term trends are often clearly visible. However, considering costs we have to move from the Binomial, Poisson or Negative Binomial error structures that appear to be reasonable for counts to alternative distributions such as Normal, Gamma or Inverse-Gaussian. Such a change can easily be handled in the classical GLM approach.

Modifications in law or in regulation more rapidly affect morbidity and health expenses than mortality. This makes long-term predictions very risky, or even meaningless. The present chapter discusses different non-linear models for projecting average yearly claim costs in health insurance

over the next few years. We study a data set issued by the German Federal Financial Supervisory Authority (BaFin), which describes average annual medical inpatient costs observed for ages 20 to 80 during calendar years 1995 to 2011. Short-term predictions are created with the help of a bilinear decomposition with log or identity link function. Different generalized bilinear models are fitted to the data. The appropriate specification for the response considered is selected by splitting the data basis into a training set (starting in 1995) and a validation set gathering the last years comprised in the data basis. After having selected the optimal model for the validation set, we perform a short-term projection of health insurance claims and explain its potential use for actuaries.

The chapter is organized as follows. Section 3.2 describes the general modeling approach. A case study with German health insurance data is given in Section 3.3. The final Section 3.4 discusses the result and briefly concludes.

3.2 Bilinear modeling

3.2.1 Model specification

Lee and Carter (1992) specified the bilinear form

$$\alpha_x + \beta_x \kappa_t \tag{3.2.1}$$

for the force of mortality on the log-scale. Please note that in the present chapter α_x relates to the age-parameter of the Lee-Carter model to comply with the standard notation in that context. This is not to be confused with the indexing factors $\alpha_x^{(k)}$ used in the preceding chapter. The specification (3.2.1) differs structurally from parametric models given that the dependence on age is nonparametric, and represented by the sequences of α_x's and β_x's. Interpretation of the parameters is quite simple: α_x is the general shape of the log-mortality schedule and the actual forces of mortality change according to an overall mortality index κ_t modulated by an age response β_x. The parameter β_x represents the age-specific patterns of mortality change. It indicates the sensitivity of the logarithm of the force of mortality at age x to variations in the time index κ_t. The specification (3.2.1) implies that the modelled death rates are perfectly correlated across ages, which is the strength but also the weakness of the approach. As pointed out by Lee (2000), the rates of decline at different ages are given by $\beta_x(\kappa_t - \kappa_{t-1})$ so that they always maintain the same ratio to one another over time.

The decomposition (3.2.1) appears to be very appealing for morbidity rates as well as for various quantities appearing in health insurance, such as age-specific expected claim frequencies or severities. The estimated α_x's exhibit the typical shape of the quantity under study (directly or on the log scale). Generally, we get relatively high values around birth, a decrease at infant ages, the accident hump, and finally the increase at adult ages with an ultimately concave behavior for quantities in line with mortality pattern. Besides the average structure captured by the α_x, the κ_t accounts for time trends that may be due to improvements in longevity, medical inflation, etc. The general time effect κ_t is adapted to each particular age by means of the factor β_x.

In Section 3.3 we explain how to apply (3.2.1) to model health insurance expenses.

3.2.2 Identifiability

In (3.2.1), the α_x parameters can only be identified up to an additive constant, the β_x parameters can only be identified up to a multiplicative constant, and the κ_t parameters can only be identified up

to a linear transformation. Precisely, if we replace β_x with $c\beta_x$ and κ_t with $\frac{\kappa_t}{c}$ for any $c \neq 0$ or if we replace α_x with $\alpha_x - c\beta_x$ and κ_t with $\kappa_t + c$ for any c, we obtain the same values for the death rates. This means that we cannot distinguish between the two parametrizations. A pair of additional constraints are required on the parameters for estimation to circumvent this problem. To some extent, the choice of the constraints is a subjective one, although some choices are more natural than others. In the literature, the parameters in (3.2.1) are usually subject to the constraints

$$\sum_t \kappa_t = 0 \text{ and } \sum_x \beta_x = 1 \tag{3.2.2}$$

ensuring model identification. Under this normalization, β_x is the proportion of change in the overall experience attributable to age x (on the log scale).

The lack of identifiability of (3.2.1) is only a minor issue. It just means that the likelihood associated with the model has an infinite number of equivalent maxima, each of which would produce identical forecasts. Adopting the constraints (3.2.2) consists in picking one of these equivalent maxima. The important point is that the choice of constraints has no impact on the quality of the fit or on forecasts.

3.2.3 Estimation

Regression models treating age and calendar time as factors are generally used to extract the α_x, β_x and κ_t from the available statistics. The products $\beta_x\kappa_t$ make them nonlinear so that standard GLM packages cannot be used. Different distributional assumptions have been proposed so far, including least-squares (Lee and Carter, 1992), Poisson (Brouhns et al., 2002a,b), Binomial (Cossette et al., 2007), and Negative Binomial (Delwarde et al., 2007) loss functions. The maximum likelihood estimates are easily found using iterative algorithms, that appear to converge very rapidly. The `gnm` package of the statistical software R (Turner and Firth, 2007) can be used to fit generalized nonlinear models with a score of the form (3.2.1), avoiding to develop such algorithms case by case. The numerical illustrations proposed in this paper are performed with `gnm`.

The specifications listed above remain relevant for morbidity rates or claim frequencies in health insurance, as these quantities are still based on event counts. However, for average severities, discrete distributions are less appealing and the actuary could consider Normal, Gamma or Inverse Gaussian specification, instead.

3.2.4 Forecast

An important aspect of the decomposition (3.2.1) is that the time factor κ_t is intrinsically viewed as a stochastic process. Box-Jenkins techniques are then used to estimate and forecast the time factor within an ARIMA time series model. These forecasts in turn yield projected age-specific mortality or morbidity rates, as well as severities, on which the calculation of premiums and reserve can be based.

3.3 Case study

3.3.1 Data description

The data set used was issued by the German Federal Financial Supervisory Authority (BaFin). It covers the period $1995-2011$. Henceforth, the response $S_x(t)$ is indexed by attained age x and

calendar time t. It describes the average yearly medical inpatient cost for year $t = 1995, \ldots, 2011$ at age $x = 20, 21, \ldots, 80$. The observed $s_x(t)$ are displayed in Figure 3.1. The shape of the data is similar to a mortality surface, with the increase at young adult ages paralleling the accident hump followed by a linear increase after age 40 similar to the Gompertz part of the mortality schedule, before an ultimate concave behavior at oldest ages. Surprisingly, the accident hump temporarily vanishes in the calendar years 1999-2001. The effect of medical inflation is also clearly visible, causing an increase in yearly health expenses as time passes.

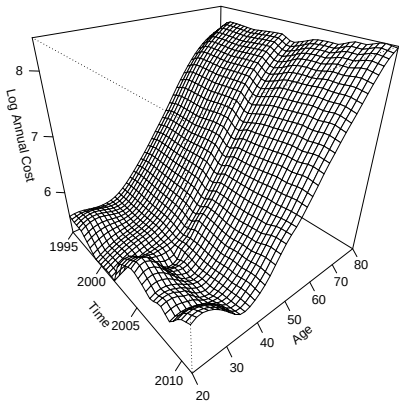

Figure 3.1: Observed inpatient costs $s_x(t)$ for German males, $t = 1995, \ldots, 2011$, $x = 20, 21, \ldots, 80$, log scale. Source: BaFin (2012).

The procedure adopted by BaFin for creating the data tables can be summarized as follows. BaFin requires data from all companies, each year, according to different tariffs and types of benefits. The collected data are crude. For each specific table (e.g. inpatient, double room), BaFin smooths the collected crude data with Whittaker-Henderson (with possibly different smoothing parameters for different age groups). There is no further adjustment of the data (except the smoothing).

In this chapter, we only consider data for German males. Data for women have been excluded for the following reason. Due to the 2004 European directive on "Equal Treatment in Goods and Services", 2004/113/EC, pregnancy costs were excluded starting in the BaFin data report in 2007, inducing a break in the data. Our analysis therefore only focuses on male data.

Figure 3.2 displays the number of policies according to policyholder's age for each calendar year 2007 to 2011. This information was missing for years up to 2006. Apart from a moderate aging effect, the available age-specific numbers appear to be relatively stable over time. Therefore, unavailable volumes for specific years were inferred from observed volumes. Specifically, the age structure has been supposed to be constant over $1995 - 2006$, as suggested by available data.

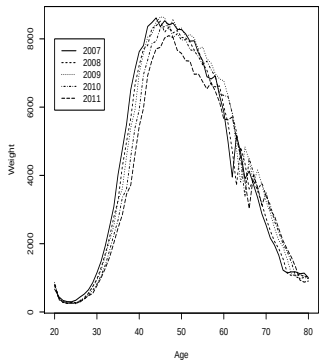

Figure 3.2: Number of policies according to policyholder's age for each calendar year 2007 to 2011. Source: BaFin (2012).

3.3.2 The Rusam method

Let us briefly explain how the expected claim costs $\mathrm{E}[S_x(t)]$ are modelled on the German market. In German private health insurance, the average yearly claim costs are usually decomposed into

$$\mathrm{E}[S_x(t)] = \gamma_x(t)\mathrm{E}[S_{x_0}(t)],$$

i.e. the expected amount at age x at time t is split into a product of an expected amount at some reference age x_0 modulated by the age pattern $\gamma_x(t)$. The profile (or age profile) of the annual medical cost data is thus defined by

$$\gamma_x(t) = \frac{\mathrm{E}[S_x(t)]}{\mathrm{E}[S_{x_0}(t)]}$$

for a fixed reference age x_0, e.g. $x_0 = 40$.

German private health insurance companies assume that $\gamma_x(t)$ is approximately constant for at least a short period of time, i.e. $\gamma_x(t) = \gamma_x(t_0)$ for all t in a small interval around t_0. By defining $\beta_x^{t_0} := \gamma_x(t_0)$ and $\kappa_t^{x_0} := \mathrm{E}[S_{x_0}(t)]$, we obtain the model

$$\mathrm{E}[S_x(t)] = \beta_x^{t_0}\kappa_t^{x_0}. \tag{3.3.1}$$

The future expected annual medical costs $\kappa_t^{x_0}$ at reference age x_0 are estimated by linear regression, based on the assumption that

$$\kappa_t^{x_0} = \kappa_{t_0}^{x_0} + \theta(t - t_0) \tag{3.3.2}$$

for some real number θ.

The assumption of time constant profiles and perfectly linear time trends are commonly referred to as Rusam method on the German market. However, Figure 3.1 shows an erratic trend pattern and changes in the profiles with respect to calendar time. Therefore, more sophisticated approaches are needed. In the present chapter we try a bilinear approach together with ARIMA time series modeling.

3.3.3 Modeling approach

In this chapter, we relax the linear trend assumption (3.3.2) underlying the Rusam approach, and we replace (3.3.1) with the specifications

$$\mathrm{E}[S_x(t)] = \alpha_x + \beta_x \kappa_t \tag{3.3.3}$$

and

$$\mathrm{E}[S_x(t)] = \exp(\alpha_x + \beta_x \kappa_t) \tag{3.3.4}$$

involving the bilinear score (3.2.1) commonly used in mortality studies with an identity or a log link function, respectively. Similar to Lee and Carter (1992), the decomposition into a age-specific pattern α_x and a time pattern κ_t (trend) modulated by an age response β_x is easy to interpret.

While $\beta_x^{t_0}$ and $\kappa_x^{t_0}$ in (3.3.2) refer to a specific age x_0 and a specific year t_0, the variables β_x and κ_t in (3.3.3) and (3.3.4) are not bound to that restriction. This fact looks negligible, but it is indeed significant when it comes to estimation. Instead of focussing on a specific age and year, the unspecific definition of β_x and κ_t in (3.3.3) and (3.3.4) reflects the aim to match the data over the full time and age span.

We estimate age profiles α_x and β_x in combination with a time profile κ_t based on all past data. The κ_t are then projected with an appropriate time series model. In contrast to the deterministic trend model (3.3.2), the time series approach for κ_t better captures empirically observed patterns and moreover allows to quantify the forecasting uncertainty.

In our case study we consider four different models. All models share similarity with the log-bilinear structure (3.2.1) underlying the approach of Lee and Carter (1992). However, we assume continuous distributions for the response variable because we do not have to deal with event counts but average claim severities instead. The models are from the class of generalized nonlinear models in the sense that:

- The response is generated from a distribution of the exponential dispersion family.
- We specify a link function (logarithm or identity function).
- The score is not a linear function of the unknown parameters.

3.3.4 Model comparisons

Henceforth, we consider the following specifications to estimate the parameters α_x, β_x and κ_t appearing in (3.3.3)-(3.3.4):

M0: Setting α_x equal to zero, the first model is in line with the Rusam method (3.3.1) with $S_x(t)$ Normally distributed with mean $\beta_x \kappa_t$ and constant variance σ^2, i.e.

$$S_x(t) \sim \mathcal{N}or(\beta_x \kappa_t, \sigma^2).$$

For this model, we only normalize the β_x, i.e. we impose the second constraint of (3.2.2). Different from the Rusam method, here β_x and κ_t do not refer to a specific year or age. Moreover, for κ_t we allow departures from the linear trend (3.3.2).

M1: This model is based on (3.3.3) with $S_x(t)$ Normally distributed, i.e.

$$S_x(t) \sim \mathcal{N}or(\alpha_x + \beta_x \kappa_t, \sigma^2).$$

This model is close to model M0 except that the set of α_x captures an average level over the observation period.

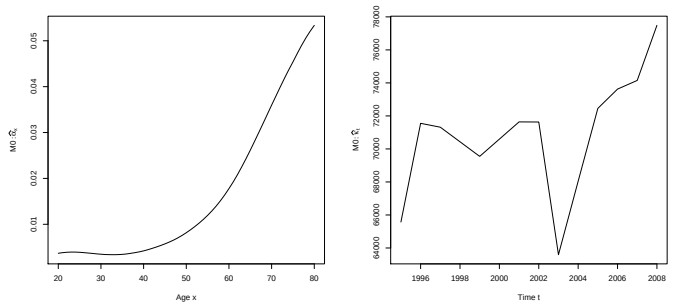

Figure 3.3: Estimated effects for model M0.

M2: This model is based on (3.3.4) with $S_x(t)$ Gamma distributed with mean $\exp(\alpha_x + \beta_x \kappa_t)$, i.e.

$$S_x(t) \sim \mathscr{G}am\Big(\exp(\alpha_x + \beta_x \kappa_t), \tau\Big).$$

This model differs from M1-M2 by the distribution assumption (Gamma vs Normal or Log-Normal) as well as by the link function (log-link vs identity link).

M3: The last model is based on (3.3.4) with $Y_x(t)$ Inverse Gaussian, i.e.

$$S_x(t) \sim \mathscr{IG}au\Big(\exp(\alpha_x + \beta_x \kappa_t), \tau\Big).$$

This model thus differs from M3 by a different distributional assumption for the response variable (Inverse Gaussian vs Gamma) but not by the link function (log-link in both cases).

The models M2 and M3 both use a log link function, which is common in mortality modeling. Different from mortality modeling, we are not modeling claim counts here but claim amounts.

3.3.5 Model selection

We split the available data set into a training set (calendar years 1995-2008) and a validation set (calendar years 2009-2011). The optimal model is selected on the basis of the accuracy in prediction of the response for the final 3-year period of the data set, in line with the actuarial applications we have in mind.

Figures 3.3-3.4 display the estimated effects for each model M0 to M3. Apart from model M2, the estimated effects have a similar pattern across ages or time. Notice the estimated change of sign for β_x in model M2, from negative at younger ages to positive at older ages. As the estimated time index κ_t is increasing, this means that average health costs tend to decrease over time at younger ages whereas they increase, as expected, at older ages. This is in contrast to the other models, which suggest increasing average health costs at all ages. Considering models M0-M1 on the one hand and model M3 on the other hand, the shape of the estimated time-sensitivities β_x appears to be reversed. For all models, the data relating to calendar year 2003 looks peculiar, as indicated by the outlying value of $\widehat{\kappa}_{2003}$. This causes a marked break in the series of the estimated time indices.

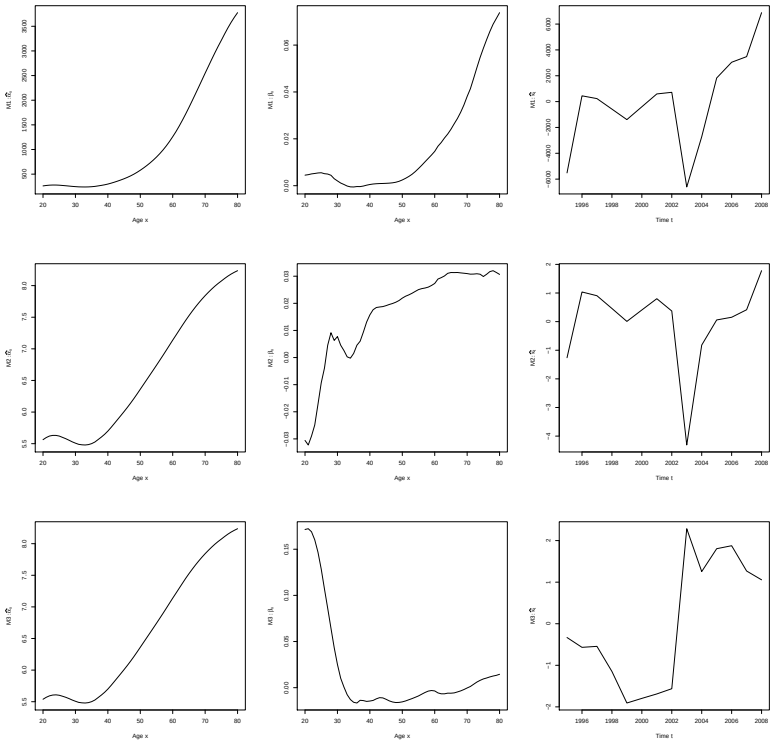

Figure 3.4: Estimated effects for models M1 (top)-M3 (bottom).

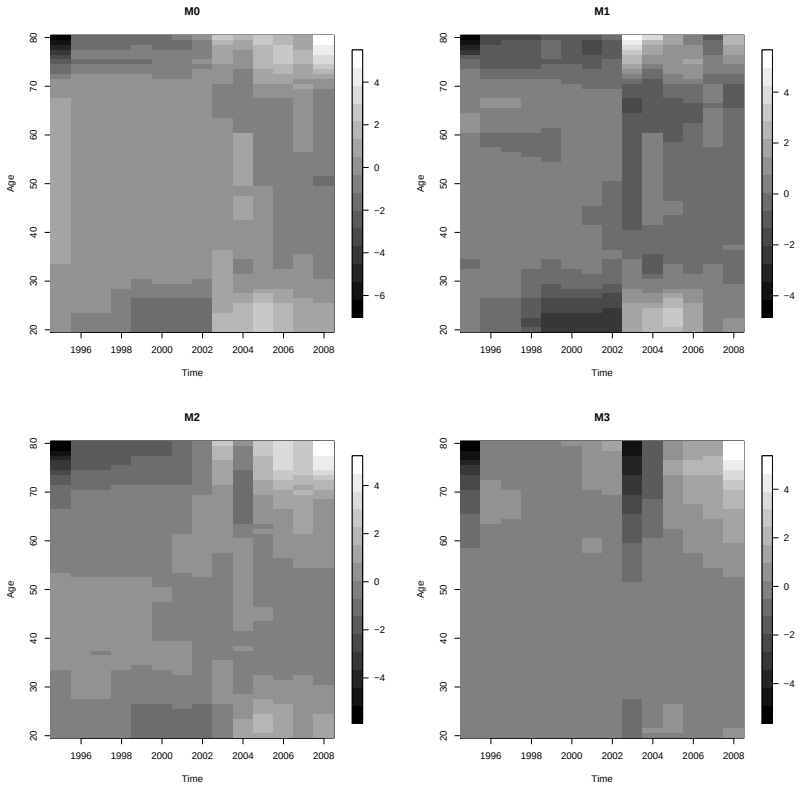

Figure 3.5: Image plots for standardized residuals r_{xt} corresponding to models M0-M3.

Since we work in a regression framework, it is essential to inspect the residuals. Model performance is assessed in terms of the randomness of the residuals. Here, we use Pearson residuals for all models M0-M3, computed from $r_{xt} = s_{xt} - \widehat{s}_{xt}$ suitably standardized. If the residuals r_{xt} exhibit some regular pattern, this means that the model is not able to describe all of the phenomena appropriately. In practice, looking at $(x,t) \mapsto r_{xt}$, and discovering no structure in those graphs ensures that the time trends have been correctly captured by the model.

Figures 3.5-3.6 display the residuals obtained for models M0-M3 as a function of both age and calendar time. Residuals are first inspected with the help of maps in Figure 3.5. The structure in the residuals can be attributed to the preliminary smoothing procedure implemented by BaFin. This induces similar values for neighboring cells. This is especially the case for M0. Residuals sometimes assume large positive or negative values near data boundaries but remain generally in the interval $[-2,2]$. This is confirmed on Figure 3.6.

In order to select the best model, we now compare the projections obtained for the last three calendar years 2009-2011 serving as validation set. To this end, the estimated κ_t's are viewed as a realization of a time series that is modelled using the classical autoregressive integrated moving

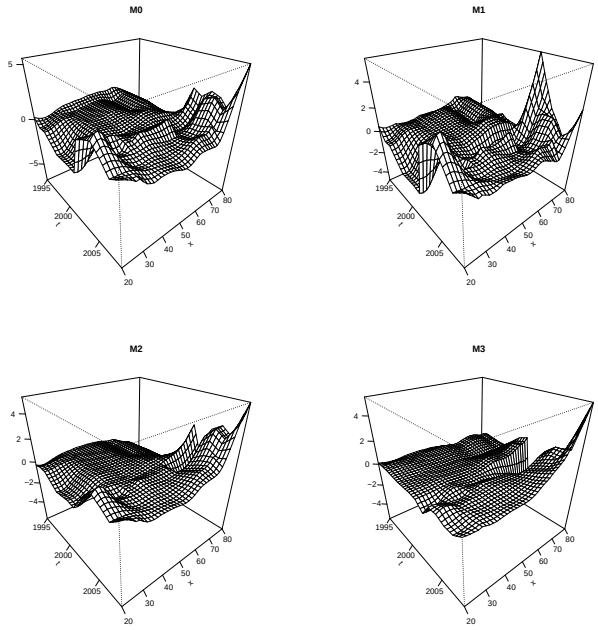

Figure 3.6: Standardized residuals r_{xt} for models M0-M3.

average (ARIMA) models. Such models explain the dynamics of a time series by its history and by contemporaneous and past shocks. In line with the standard Lee-Carter approach, estimated κ_ts are projected using an ARIMA (0,1,0) time series model. The results are visible on Figure 3.7. We can see there that the relative differences obtained with M1 outperform the three other models, staying in the range $\pm 5\%$ for 2009 and achieving relative errors less than 10% for 2010 and 2011, except at younger ages. The Rusam method performs particularly weak in the age range 20 to 30. In that age range Figure 1 shows a time pattern that is highly non-linear and clearly different from the time pattern at reference age x0 = 40. Both effects are contrary to the assumptions underlying Rusam's method. At higher ages, the time pattern becomes more linear and uniform and the difference between methods M0-M3 and Rusam's method is less profound.

The quality of the prediction is measured with the square root of the mean squared error (RMSE) as well as with the mean absolute error (MAE). Here, the mean squared error is obtained from averaging the squared difference between the observations s_{xt} for $t \in \{2009, 2010, 2011\}$ and their prediction $\widehat{s}_{xt}$ obtained from the α_x and β_x estimated on the training set combined with the κ_t projected to $t \geq 2009$ with the help of the random walk with drift model fitted to the time index of the training period. The mean absolute error is obtained similarly, by taking the absolute differences between observed s_{xt} for $t \in \{2009, 2010, 2011\}$ and their prediction $\widehat{s}_{xt}$. The results are reported in Table 3.1. Based on the performances on the validation set, model M1 appears to be the best one. Note that the weak performance of Rusam's method at young ages has a rather small impact here since the claim costs are very small at young ages.

	Rusam	M0	M1	M2	M3
Year	Mean Absolute Error (MAE)				
2009	56.07801	44.51098	24.26994	50.66862	104.9483
2010	133.43555	71.68413	51.74541	111.81426	175.1147
2011	110.58424	85.13883	67.72094	90.35197	145.1686
Year	Root Mean Squared Error (RMSE)				
2009	76.96069	58.72724	34.49802	99.74894	178.2142
2010	189.67460	130.86420	79.59536	195.99804	280.8031
2011	149.84950	110.01762	88.29558	150.69008	227.3208

Table 3.1: Root Mean Squared Error and Mean Absolute Error for the validation set.

This observation is confirmed by an economic analysis of the prediction results. The net present value of the health expenditure for high ages (starting at age 61) of the observations $\sum_{x>60} s_{xt} \cdot 1.01^{-(x-60)}$ is compared with the value of their prediction $\sum_{x>60} \widehat{s}_{xt} \cdot 1.01^{-(x-60)}$ for $t \in \{2009, 2010, 2011\}$. The yearly discount rate is set to 1%. The results are shown in Table 3.2. The results reveal that the prediction of the models M2 and M3 clearly underestimate the health expenses. The weak performance of Rusam's method at young ages is here dampened by the small weight that young ages have on the total costs. If we focused on young ages only, the mismatch would increase significantly. Again, the model M1 tends to be the best choice. In particular, the net present value of the prediction only slightly overestimates the net present value of the observations. This is advantageous for risk management purposes.

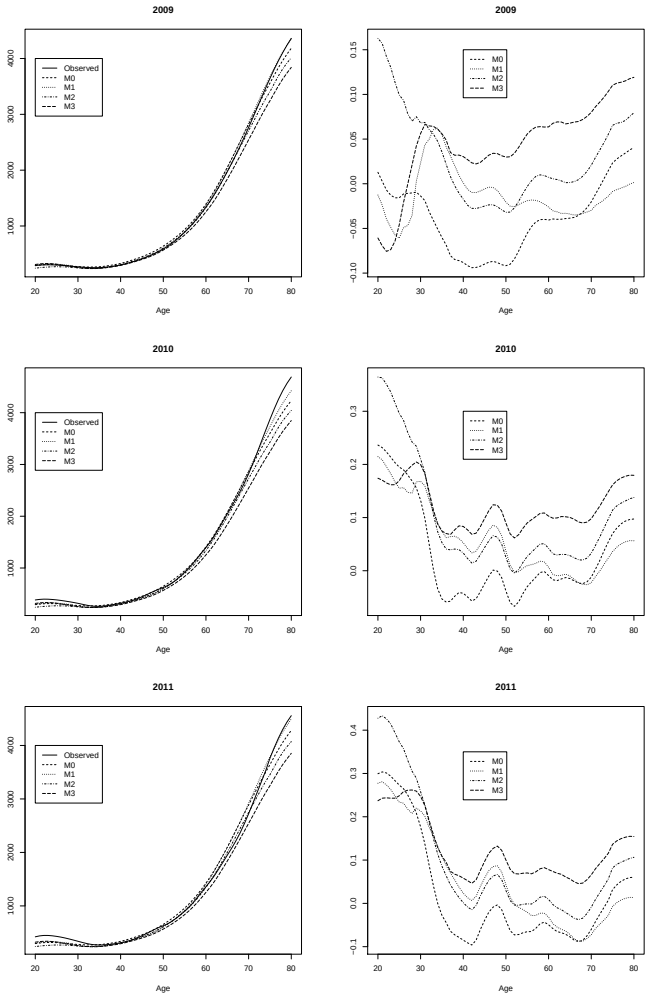

Figure 3.7: Comparisons of actual s_{xt} vs predictions obtained from the training set for models M0-M3 with relative differences $(s_{xt} - \widehat{s}_{xt})/s_{xt}$.

3.3.6 Projection

Let us now project the average yearly health claims three years beyond the end of the observation period, thus for calendar years 2012 to 2014, with the help of the optimal model M1 that has been selected based on the validation set.

To this end, we first fit this model to the entire observation period 1995-2011. The results are given in Figure 3.8. The estimated age effect α_x and β_x are similar to those displayed on Figure 3.4

	Observation	Rusam	M0	M1	M2	M3
Year	Net Present Value of Health Expenses					
2009	51.066	51,066	51,042	51,989	48,993	46,278
2010	53.541	53,541	51,638	52,705	49,349	46,298
2011	51.703	51,703	52,235	53,421	49,707	46,318

Table 3.2: Net Present Value of Health Expenses for the validation set.

obtained based on the training set 1995-2008. The time indices are now supplemented with three values, for calendar years 2009-2011 that are now included in the analysis. A linear trend is clearly visible for the $\widehat{\kappa}_t$.

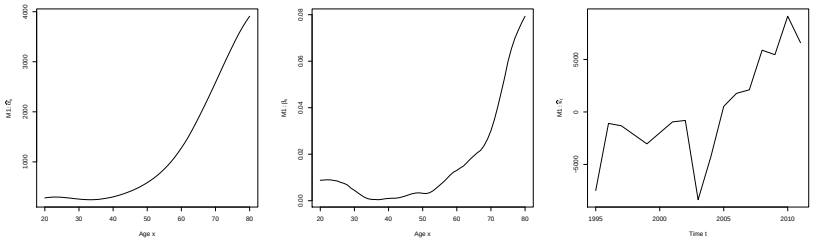

Figure 3.8: Estimated effects for model M1 fitted to the entire observation period 1995-2011.

Plugging the projected time index to 2014 in the bilinear decomposition provides the actuary with the forecast of health expenditures over the next 3 years. However, this point prediction, while interesting, reveals nothing about the uncertainty attached to the future health costs. In forecasting, it is important to provide information about the error affecting the quantities of interest. In that respect, prediction intervals are particularly useful. In the current application, it is impossible to derive the relevant prediction intervals analytically. The reason for this is that two very different sources of uncertainty have to be combined: sampling errors in the parameters α_x, β_x, and κ_t, and forecast errors in the projected κ_t's. An additional complication is that the measures of interest $\mathrm{E}[S_{xt}]$ are non-linear functions of the parameters α_x, β_x, and κ_t and of the ARIMA parameters. This is why bootstrap procedures are used.

The key idea behind the bootstrap is to resample from the original data (either directly or via a fitted model) in order to create replicate data sets, from which the variability of the quantities of interest can be assessed. Because this approach involves repeating the original data analysis procedure with many replicate sets of data, it is sometimes called a computer-intensive method. Bootstrap techniques are particularly useful when, as in our problem, theoretical calculation with the fitted model is too complex.

If we ignore the other sources of errors, then the confidence bounds on future κ_t's can be used to calculate prediction intervals. For mortality studies, we know from Lee and Carter (1992, Appendix B), that prediction intervals based on κ_t alone are a reasonable approximation only for forecast

horizons of 10 to 25 years. For long-run forecasts, the error in forecasting the time index thus dominates the errors in fitting the mortality rates. If there is a particular interest in forecasting over the shorter term, as it is the case here, then we cannot make a precise analysis of the forecast errors and prediction intervals based on κ_t alone seriously understate the errors in forecasting over shorter horizons.

This is why we derive here bootstrap percentiles confidence intervals. The bootstrap procedure yields B samples of α_x, β_x, and κ_t parameters, denoted as α_x^b, β_x^b, and κ_t^b, $b = 1,2,\ldots,B$. This procedure can be carried out in several ways (Brouhns et al., 2002a, b, 2005, Koissi et al., 2006): by Monte Carlo simulation from the approximate multivariate Normal distribution of the maximum likelihood estimators $\widehat{\alpha}_x$, $\widehat{\beta}_x$, and $\widehat{\kappa}_t$, by parametric bootstrap resampling from the fitted model and then re-estimating the age and time effects, or by bootstrapping the residuals.

As reported by Renshaw and Haberman (2008), the result of Monte Carlo simulation invoking the large sample properties of the maximum likelihood estimators heavily rely on the identifiability constraints. Given that the choice of constraints is not unique and that this choice materially affects the resulting simulations, this first approach should not be used for risk assessment purposes unless there are compelling reasons for selecting a particular set of identifiability constraints. As the preliminary smoothing implemented by BaFin generates local similarities, we conduct parametric bootstrap by generating s_{xt}^b as realizations from random variables obeying the $\mathcal{N}or(\alpha_x + \beta_x\kappa_t, \sigma^2)$ distribution. The age and time effects are then re-estimated using the s_{xt}^b as data points, producing α_x^b, β_x^b, and κ_t^b, $b = 1,2,\ldots,B$.

We then estimate the time series model using the κ_t^b as data points. This yields a new set of estimated ARIMA parameters. We can then generate a projection κ_t^b for t beyond the observation period using these ARIMA parameters. In the random walk with drift model, it is enough to add to the last available $\widehat{\kappa}_t^b$ as many times the drift estimated from the κ_t^b data points as needed to reach the projection horizon.

The first step is meant to take into account the uncertainty in the parameters α_x's, β_x's and κ_t's. The second step deals with the fact that the uncertainty in the ARIMA parameters depends on the uncertainty in the α_x's, β_x's and κ_t's parameters. The third step ensures that the uncertainty of the forecasted κ_t's depends on the uncertainty of the ARIMA parameters themselves.

This yields B realizations α_x^b, β_x^b, κ_t^b and projected κ_t^b on the basis of which we can compute the measure of interest $\widehat{s}_{xt}^b$ corresponding to $\mathrm{E}[S_{xt}]$. Assume that B bootstrap estimates $\widehat{s}_{xt}^b$, $b = 1,2,\ldots,B$, have been computed. The $(1-2\alpha)$ percentile interval for $\mathrm{E}[S_{xt}]$ is given by $(\widehat{s}_{xt}^{b(\alpha)}, \widehat{s}_{xt}^{b(1-\alpha)})$, where $\widehat{s}_{xt}^{b(\zeta)}$ is the $100\times\zeta$th empirical percentile of the bootstrapped values for $\widehat{s}_{xt}^b$, which is equal to the $(B\times\zeta)$th value in the ordered list of replications $\widehat{s}_{xt}^b$, $b = 1,2,\ldots,B$. For instance, in the case of $B = 1,000$ bootstrap samples, the 0.95th and the 0.05th empirical percentiles are, respectively, the 950th and 50th numbers in the increasing ordered list of 1,000 replications of $\mathrm{E}[S_{xt}]$.

The predictions of $\mathrm{E}[S_{xt}]$ for calendar years 2012-2014 are displayed in Figure 3.9, together with 90% prediction intervals and relative accuracies at the same confidence level (appearing colored in gray) obtained by parametric bootstrap (with $B = 1,000$). We can see there that the relative accuracy on the expected health insurance claims is generally in the range $\pm 10\%$ at most ages.

In order to figure out the predictive distribution of S_{xt}, we produce analogs to the longevity fan charts proposed by Dowd et al. (2010) based on parametric bootstrap (with $B = 1,000$). The result is shown in Figures 3.10-3.12. These charts depict some central projection of the forecasted variable S_{xt}, together with bounds around this showing the probabilities that the variable will lie within specified ranges. The difference with the results displayed in Figure 3.9 if that the time index is now projected to 2012-2014 allowing for some departures from the average trajectory, and that

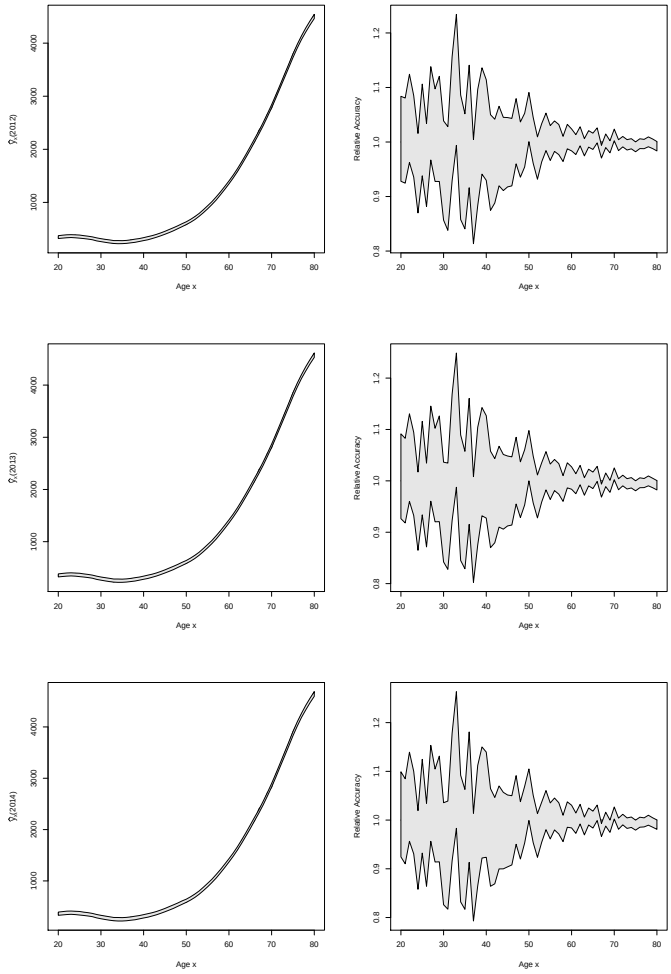

Figure 3.9: Predictions of s_{xt} with model M1 fitted to the entire observation period 1995-2011 for calendar years 2012 (top)-2014 (bottom) with 90% bootstrap prediction intervals, and corresponding relative accuracies.

realizations of future S_{xt} are finally simulated from the Gaussian distribution. The chart in Figures 3.10-3.12 shows the central 10% prediction interval with the heaviest shading surrounded by the 20%, 30%, ..., 90% prediction intervals with progressively lighter shading. The shading becomes stronger as the prediction interval narrows. We can therefore interpret the degree of shading as reflecting the likelihood of the outcome: the darker is the shading, the more likely is the outcome.

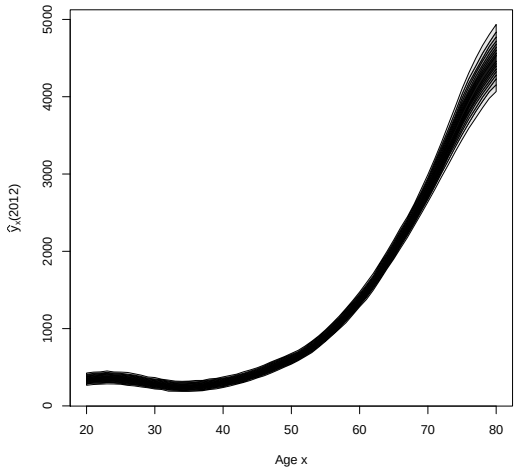

Figure 3.10: Fan chart for $S_x(2012)$.

The fan in Figures 3.10-3.12 consists of 9 grey bands of varying intensity. The upper and lower boundaries correspond to paths of the forecast 95% and 5% quantiles, and the inner edges of the bands in the fan correspond to the 10%, 15%, ..., 90% quantiles. The darkest band in the middle is bounded by the 45% and 55% quantiles. Note that the quantiles are calculated for each year in isolation. The fan charts in Figures 3.10-3.12 show that short-term departures from the central trend remain limited, except at older ages where the spread is more pronounced.

To end with, let us make a last remark. Future $\mathrm{E}[S_{xt}]$ have been obtained using extrapolated κ_t's and fixed α_x's and β_x's. In this case, the jump-off values (i.e., $\mathrm{E}[S_{xt}]$ in the last year of the fitting period or jump-off year) are fitted rates. The Lee-Carter method has been criticized by Bell (1997) for the fact that a discontinuity is possible between the observed mortality rates and life expectancies for the jump-off year and the forecast values for the first year of the forecast period. The bias arising from this discontinuity would then persist throughout the forecast. The same comments apply to our setting, and the forecast could start with the last observed s_{xt} or their average over the end of the observation period.

3.4 Remark on overfitting

Overfitting occurs when the model is excessively complex, such as having too many parameters relative to the number of observations. An overfitted model has poor predictive performance and it overreacts to minor fluctuations in the training data.

To avoid overfitting we have divided our dataset into a training set and a validation set. Our model performs well both on the training set and on the validation one and the predictions are well reproducing the data for the period 2009-2011.

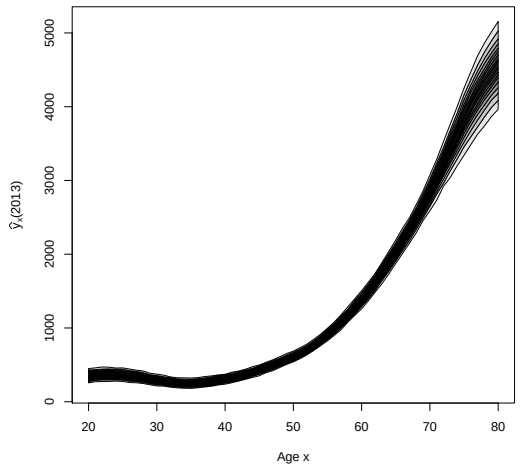

Figure 3.11: Fan chart for $S_x(2013)$.

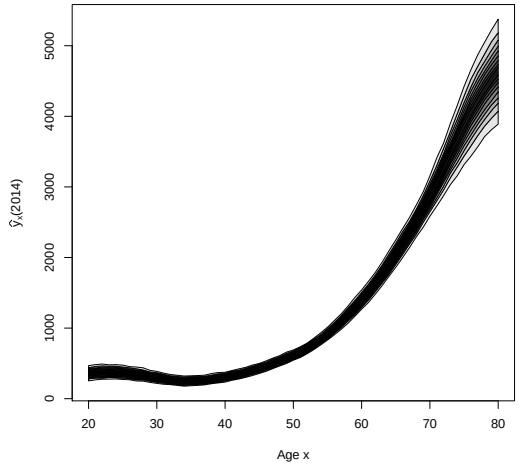

Figure 3.12: Fan chart for $S_x(2014)$.

We have provided on Figure 3.13 and 3.14 long term predictions of our model, i.e. period 2020-2040 and the projections conform with our expectations.

As pointed out in Delwarde et al. (2007), it is possible to reduce model complexity by imposing smoothness on the estimated the estimated β_x's in the Lee Carter and Poisson log-bilinear models for mortality projection. To this end, penalized least-squares or penalized log-likelihood maximization is performed. To smooth the estimated β_xs, the following objective function could be used:

$$\sum_{x=x_{\min}}^{x_{\max}} \sum_{t=t_{\min}}^{t_{\max}} (\ln\hat{S}_x(t) - \alpha_x - \beta_x \kappa_t)^2 + \beta' P_\beta \beta$$

where

$$P_\beta = \pi_\beta \delta' \delta$$

with

$$\delta = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix}$$

π_β is the smoothing parameter. The choice of the optimal π_β will be based on the observed data, using cross validation.

Note that in Delwarde et al. (2007) the objective function can therefore be seen as a compromise between goodness of fit (first term) and smoothness of the β_xs (second term). The penalty involves the sum of the squared second-order differences of the β_xs, that is, the sum of the square of the second differences $\beta_{x+2} - 2\beta_{x+1} + \beta_x$s.

Some other solutions have been proposed in case of severely fluctuating predicted age-specific mortality in the Lee-Carter model. We refer to Zhao (2012) which introduces a new modified Lee-Carter model for analysing short-base-period mortality data, and approximates the unknown parameters in the modified model by linearized cubic splines and other additive functions. We also refer to Hainaut (2018) and Hong (2020) which propose hybrid methods, based on Lee-Carter with ANN (artificial neural network) or RF (random forest) for mortality projection.

3.5 Conclusion

In German private health insurance, the so-called Rusam method is the traditional way to create short-term predictions, assuming the time trend to be perfectly linear and the age profile to be time-constant. However, the BaFin data shows erratic time trends and non-constant age profiles with respect to calendar time.

Inspired by the literature on modeling and projection of mortality rates, in this chapter we performed short-term predictions of health insurance claim costs by using a bilinear approach and ARIMA time series modeling. The bilinear approach better captures the non-constant age profiles, and the ARIMA time series tools allow to capture the erratic patterns in the time trend factor. While in mortality modeling a log link function is commonly used, the analysis performed here suggests that for modeling health expenses the identify link function works better. By calculating individual net present values, we could see the economic relevance of choosing a good model.

Since the empirically observed time trend is quite volatile, confidence interval estimates rather than point estimates should be used for predictions. Here, bootstrap techniques were applied for creating prediction intervals, yielding very reasonable results.

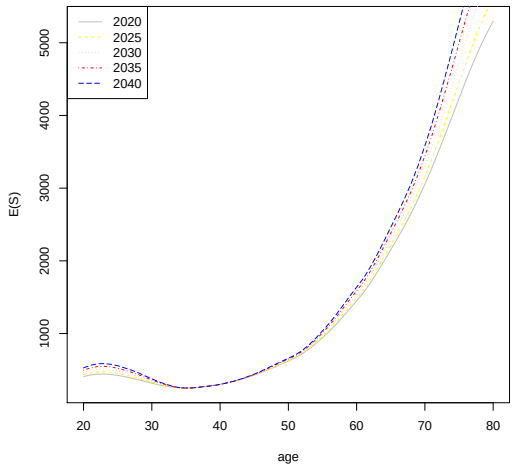

Figure 3.13: Model long-term forecasts from 2020 until 2040

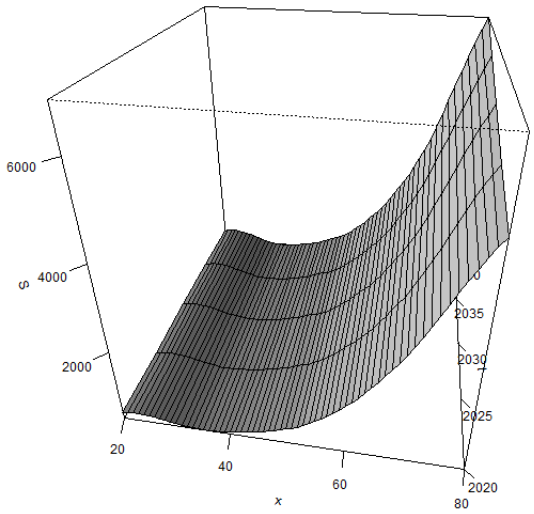

Figure 3.14: Model long-term forecasts from 2020 until 2040, 3D representation

References

- BaFin or Federal Financial Supervisory Authority (2012).
  Wahrscheinlichkeitstafeln in der privaten Krankenversicherung 2012.
  Available from http://www.bafin.de/SharedDocs/Veroeffentlichungen/DE/Statistik/st_wahrscheinlichkeitstafeln_pkv_2012.html
- Bell, W. (1997). Comparing and assessing time series methods for forecasting age-specific fertility and mortality rates. Journal of Official Statistics 13, 279-303.
- Brouhns, N., Denuit, M. & Van Keilegom, I. (2005). Bootstrapping the Poisson log-bilinear model for mortality projection. Scandinavian Actuarial Journal 2005, 212-224.
- Brouhns, N., Denuit, M. & Vermunt, J.K. (2002a). A Poisson log-bilinear approach to the construction of projected life tables. Insurance: Mathematics and Economics 31, 373-393.
- Brouhns, N., Denuit, M. & Vermunt, J.K. (2002b). Measuring the longevity risk in mortality projections. Bulletin of the Swiss Association of Actuaries 2002, 105-130.
- Christiansen, M., Denuit, M. & Lazar, D. (2012). The Solvency II square-root formula for systematic biometric risk. Insurance: Mathematics and Economics 50, 257-265.
- Cossette, H., Delwarde, A., Denuit, M., Guillot, F. & Marceau, E. (2007). Pension plan valuation and dynamic mortality tables. North Americal Actuarial Journal 11, 1-34.
- Delwarde, A., Denuit, M. & Partrat, Ch. (2007). Negative Binomial version of the Lee-Carter model for mortality forecasting. Applied Stochastic Models in Business and Industry 23, 385-401.
- Denuit, M., Dhaene, J., Hanbali, H., Lucas, N. & Trufin, J. (2017) Updating mechanism for lifelong insurance contracts subject to to medical inflation. European Actuarial Journal 7, 133-163.
- Dhaene, J., Godecharle, E., Antonio, K., Denuit, M. & Hanbali, H. (2017). Lifelong health insurance covers with surrender value: Updating mechanisms in the presence of medical inflation. ASTIN Bulletin, 47(3), 803-836.
- Dowd, K., Blake, D. & Cairns, A. (2010). Facing up to uncertain life expectancy: The longevity fan charts. Demography 47, 67-78.
- Hainaut, D. (2018) A neural-network analyzer for mortality forecast. ASTIN Bulletin: The Journal of the IAA 48(2), 481-508.
- Hong, W., Yap, J., Selvachandran, G. & Thong, P. (2020) Forecasting mortality rates using hybrid LeeâĂŞCarter model, artificial neural network and random forest. Complex & Intelligent Systems, 1-27.
- Koissi, M., Shapiro, A.F. & Hognas, G. (2006). Evaluating and extending the Lee-Carter model for mortality forecasting: Bootstrap confidence intervals. Insurance: Mathematics and Economics 38, 1-20.
- Lee, R. (2000). The Lee-Carter method of forecasting mortality, with various extensions and applications. North American Actuarial Journal 4, 80-93.
- Lee, R. & Carter, L. (1992). modeling and forecasting the time series of US mortality. Journal of the American Statistical Association 87, 659-671.

- Lee, R. & Miller, T. (2001). Evaluating the performance of the Lee-Carter approach to modeling and forecasting. Demography 38, 537-549.
- Levantesi, S. & Menzietti, M. (2012). Managing longevity and disability risks in life annuities with long term care. Insurance: Mathematics and Economics 50, 391-401.
- Renshaw, A. E. & Haberman, S. (2000). modeling the recent time trends in UK permanent health insurance recovery, mortality and claim inception transition intensities. Insurance: Mathematics and Economics 27, 365-396.
- Turner, H. & Firth, D. (2007). gnm: A package for generalized nonlinear models. R News 7, 8-12.
- Zhao, B. (2012) A modified Lee-Carter model for analysing short-base-period data. Population Studies 66(1), 39-52.

Chapter 4

Joint analysis of mortality and morbidity trends

The present chapter is based on the following paper:

- Lucas, N., Avalosse, H. Denuit, M. (2020). Hospital inpatients costs dynamics at older ages (No. 2020027). UC Louvain, Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA). Submitted to Annals of Actuarial Sciences on September 25, 2020.

4.1 Introduction

It is now well documented that if proximity-to-death is ignored, then the apparent effect of age on health care expenditures will be exaggerated. We refer the reader e.g. to Zweifel et al. (1999), Miller (2001), Zweifel et al. (2004), and Breyer and Felder (2006). This is because individuals who get close to death (who are older on average) tend to have much higher health care expenditures than those at the same age who survive. Payne et al. (2007) reviewed the literature devoted to age-based and time-to-death models for health expenditures. Time-to-death models count backward from a fixed reference point (a known date of death) and measure expenditures against this backward count. This approach thus controls the effect of longevity, and hence can offer more accurate forecasts of future expenditures.

This chapter further extends the approach proposed by Felder et al. (2010) who investigated whether time-to-death is related to health care expenditures as recorded by a sickness fund. The health care costs for a given age group of people are split into expenditures for those who die within the year (decedents) to those still living (survivors) at different time horizons. We investigate the "red herring" hypothesis, stating that attained age is of secondary importance once proximity-to-death is controlled for. See e.g. Zweifel et al. (1999). Several authors have disputed the robustness of the "red herring" findings, e.g. because health care expenditures are zero-inflated and otherwise obey a skewed distribution. This issue has been addressed by Werblow et al. (2007) by employing a two-part model separating the selection part (probability of positive health care expenditures). Here, we extend this approach by adopting a frequency-severity decomposition of total health care expenditures (HCE).

Indeed, we provide accurate modeling of dynamics in both frequency and severity components, which is of utmost importance for the insurance industry. In this chapter, total yearly costs are de-

composed into frequency and severity components to separate their effects. Precisely, the frequency component corresponds to the number of admissions to hospital whereas the severity component corresponds to the cost per stay at hospital. Such a decomposition helps to clarify the relationship between age, mortality, and morbidity among the elderly and enables more accurate expenditure forecasts by isolating medical inflation. This decomposition of total cost into frequency and severity components is in line with the proposal by Frees et al. (2011) for individual, longitudinal health expenses data except that we deal here with aggregated data so that the two-part modeling proposed by these authors is replaced with a collective compound Poisson model.

A detailed case study performed on Belgian data illustrates the modeling strategy proposed in this chapter. We used an extended data set of observed inpatients hospital care costs furnished by one of the largest sickness funds dominating the Belgian market of compulsory health insurance, i.e. the Christian Mutualities (Alliance Nationale des Mutualités Chrétiennes - Landsbond der Christelijke Mutualiteiten, henceforth referred to as ANMC), based in Brussels. It corresponds to about 4.07 million individuals, 12.5% of whom died between 2002 and 2019, with comprehensive inpatients hospital care costs data for the years 2002-2019. Those individuals make up the so-called general regime, which excludes the self-employed workers. The hospital costs comprise the part covered by the sickness fund together with out-of-pocket payments, as stated on the hospital bill. For confidentiality reasons, data have been aggregated and we use a collective compound Poisson model for the total yearly costs observed over groups of individuals cross-classified according to age, gender, calendar time, and proximity to death. Projected life tables produced by the Federal Planning Bureau based in Brussels are used to include the longevity component in the proposed modeling strategy, to combine the projections according to proximity to death in a short to medium term forecast. A comparison with the alternative model targeting total costs, not distinguishing between frequency and severity components, demonstrates the superior explanatory power of the proposed approach, revealing that total costs are mainly driven by their frequency component for the data under consideration.

This chapter provides a better understanding of the relationship between mortality and morbidity at the end of life and how this relationship might change both over time and with age at death. The contributions are as follows. We provide an explanation of the apparent stability or even decrease of the observed average yearly hospital cost over time, by showing that opposite trends act on the average number of hospital admissions and on the average cost per hospital stay. Longevity and medical inflation effects can be disentangled with the help of an appropriate frequency-severity decomposition. The chapter also demonstrates that the decedents-versus-survivors model may distort the shape of hospital expenses because the effect of attained age on both frequencies and severities sometimes differs between healthy individuals (that is, those individuals who survive several years) and those dying in the forthcoming years. We produce forecasts introducing longevity dynamics in the model and therefore linking morbidity to mortality projections.

The chapter is organized as follows. Section 4.2 describes the proposed modeling strategy. Section 4.3 is devoted to the case study based on Belgian data. The final Section 4.4 briefly discusses the results.

4.2 Modeling strategy

4.2.1 Data under consideration

In this chapter, the data have been aggregated for confidentiality reasons. Individual observations are cross-classified according to gender, attained age x, calendar year y, and proximity to death. Male and female health expenses are studied separately to capture gender-specific dynamics. Within each age-time category, individual observations are grouped according to their respective remaining lifetimes. Precisely, we distinguish among individuals aged x who are going to die in the next year, before reaching age $x+1$, from those who survive up to age $x+k-1$ but die before reaching age $x+k$, $k = 2,3,\ldots,m$, and finally those who die after age $x+m$ who form the last, open category. We thus have $m+1$ categories according to proximity to death: the decedents and m categories of survivors, depending on their times to death.

Let $L_x(y)$ be the number of individuals aged x in calendar year y covered by the sickness fund. Since we concentrate on older ages (above 70 in the case study), we can neglect new entrants and consider that the only exit cause is due to death. This is because elderly people generally stay with the same sickness fund so that arrivals and departures are negligible. Proper exposures (or exposures-to-risk, henceforth abbreviated as ER) are computed with the help of observed numbers of deaths by assuming a uniform distribution over the year (so that people die on average in the middle of the year). Precisely, the exposure is 0.5 for the decedents and 1 for the different categories of survivors.

4.2.2 Frequency-severity decomposition

Denote as $T_x(y)$ the remaining lifetime, or time to death for an individual aged x in calendar year y. This means that an individual aged x in calendar year y dies at age $x+T_x(y)$ in calendar year $y+T_x(y)$. Proximity to death is assessed with the help of categories $T_x(y) \leq 1$, $1 < T_x(y) \leq 2$, …, $m-1 < T_x(y) \leq m$, and $T_x(y) > m$. The total number $L_x(y)$ of individuals aged x in calendar year y is then split into $L_x^{T_x(y)<1}(y)$, $L_x^{j<T_x(y)\leq j+1}(y)$, $j = 1,...,m-1$ and $L_x^{T_x(y)>m}(y)$ for each category. For each proximity to death j, the total costs are then broken into two components:

- the frequency component corresponding to the total number of hospital admissions, that reflects morbidity and age rationing.
- the severity component corresponding to the cost per hospital stay, that reflects medical inflation as well as age rationing.

Precisely, the total inpatients hospital costs $S_x^{j<T_x(y)\leq j+1}(y)$ for individuals aged x in calendar year y, who survive up to age $x+j$ but die in calendar year $y+j$ before reaching age $x+j+1$, is decomposed into the sum of the respective costs $C_{x,k}^{j<T_x(y)\leq j+1}(y)$ of the $N_x^{j<T_x(y)\leq j+1}(y)$ stays at hospital for these individuals, that is,

$$S_x^{j<T_x(y)\leq j+1}(y) = \sum_{k=1}^{N_x^{j<T_x(y)\leq j+1}(y)} C_{x,k}^{j<T_x(y)\leq j+1}(y) \tag{4.2.1}$$

where $N_x^{j<T_x(y)\leq j+1}(y)$ is the number of admissions to hospital at age x in year y and $C_{x,1}^{j<T_x(y)\leq j+1}(y)$, $C_{x,2}^{j<T_x(y)\leq j+1}(y)$, … denote the associated costs. All the random variables are assumed to be independent and the severities $C_{x,1}^{j<T_x(y)\leq j+1}(y)$, $C_{x,2}^{j<T_x(y)\leq j+1}(y)$, … are assumed to be identically

distributed. Under the stated assumptions, we have

$$\mathrm{E}[S_x^{j<T_x(y)\leq j+1}(y)] = \mathrm{E}[N_x^{j<T_x(y)\leq j+1}(y)]\mathrm{E}[C_{x,1}^{j<T_x(y)\leq j+1}(y)]. \tag{4.2.2}$$

Dividing $S_x^{j<T_x(y)\leq j+1}(y)$ by the corresponding exposure gives the average total cost per unit of exposure (per person-year, thus). Similarly, dividing $N_x^{j<T_x(y)\leq j+1}(y)$ by the corresponding exposure gives the average frequency of admission (per person-year) while dividing $S_x^{j<T_x(y)\leq j+1}(y)$ by $N_x^{j<T_x(y)\leq j+1}(y)$ gives the average cost per hospital stay. These variables are denoted respectively as $\overline{S}_x^{j<T_x(y)\leq j+1}(y)$, $\overline{N}_x^{j<T_x(y)\leq j+1}(y)$ and $\overline{C}_x^{j<T_x(y)\leq j+1}(y)$ and are the responses of interest in the remainder of this chapter. Clearly, the identity

$$\overline{S}_x^{j<T_x(y)\leq j+1}(y) = \overline{N}_x^{j<T_x(y)\leq j+1}(y)\overline{C}_x^{j<T_x(y)\leq j+1}(y)$$

holds true.

4.2.3 Nonlinear regression models

Non-linear regression models are fitted to observed values $\overline{s}_x^{j<T_x(y)\leq j+1}(y)$, $\overline{n}_x^{j<T_x(y)\leq j+1}(y)$ and $\overline{c}_x^{j<T_x(y)\leq j+1}(y)$. These models are based on distributional assumptions corresponding to the characteristics of the response: event counts, average claim severities or average yearly totals. For the frequency component $\overline{N}_x^{j<T_x(y)\leq j+1}(y)$, we use Poisson regression while Gamma regression is adopted for the severity component $\overline{C}_x^{j<T_x(y)\leq j+1}(y)$. This allows us to compare the results with those produced by a Tweedie regression model run on $\overline{S}_x^{j<T_x(y)\leq j+1}(y)$.

In our case study, we consider regression models from the class of GAMs (Generalized Additive Models) in the sense that:

- the response obeys a distribution belonging to the exponential dispersion family.
- a logarithmic link function is specified.
- the score is not a linear function of the unknown parameters.

The mean value of the response is of the form $\exp(s(x,y))$ where $s(\cdot,\cdot)$ is a smooth function of age x and calendar year y. In the case study, the function $s(\cdot,\cdot)$ is estimated with the help of splines in the corresponding GAM. Of course, any other regression model could be used instead, such as zero-adjusted or zero-augmented versions of the distributions considered here, for instance.

4.2.4 Projection

Proximity-to-death $T_x(y)$ is not a variable that is known beforehand: we do not know when an individual alive at age x will die so that we cannot allocate this individual to the right category in terms of proximity to death. Having fitted a specific model for each proximity category, that is, for $T_x(y) \leq 1$, $1 < T_x(y) \leq 2$, …, $T_x(y) > m$ the outputs of these models have to be aggregated in order to make actual predictions. This can be done as explained next.

Assume that we are now in year y so that $T_x(y)$ have not been observed yet and individuals cannot be allocated to different categories in terms of proximity to death. Denote as $\mathrm{ER}_x^{T_x(y)<1}(y)$, $\mathrm{ER}_x^{j<T_x(y)\leq j+1}(y)$ and $\mathrm{ER}_x^{T_x(y)>m}(y)$ the expected risk exposures in each category $T_x(y) < 1$, $j <$

$T_x(y) \leq j+1$, $j = 1,\ldots,m-1$, and $T_x(y) > m$, respectively. The total exposure is then obtained from

$$\mathrm{ER}_x^{\mathrm{tot}}(y) \quad = \quad \mathrm{ER}_x^{T_x(y)<1}(y) + \sum_{j=1}^{m-1} \mathrm{ER}_x^{j<T_x(y)\leq j+1}(y) + \mathrm{ER}_x^{T_x(y)>m}(y).$$

We introduce the following notations for the one-year survival and death probabilities:

$$p_x(y) = \mathrm{P}[T_x(y) > 1] = 1 - q_x(y).$$

The j-year survival probability is then given by

$${}_jp_x(y) = \prod_{k=0}^{j-1} p_{x+k}(y+k).$$

Assuming a uniform distribution of death within the year, deceased individuals survive up to the middle of the year, on average, and we then obtain

$$\mathrm{ER}_x^{T_x(y)<1}(y) = L_x(y)q_x(y)\frac{1}{2}$$

For the other proximity-to-death categories, we have:

$$\mathrm{ER}_x^{j<T_x(y)\leq j+1}(y) = L_x(y)\,{}_jp_x(y)q_{x+j}(y+j), \quad j = 1,\ldots,m-1,$$

and

$$\mathrm{ER}_x^{T_x(y)>m}(y) = L_x(y)\,{}_mp_x(y).$$

Having a frequency model and a severity model including proximity to death, the expected value of the total cost $S_x(y)$ for individuals aged x in year y can then be obtained from

$$\begin{aligned}
\mathrm{E}[S_x(y)] \quad = \quad & \frac{\mathrm{ER}_x^{T_x(y)<1}(y)}{\mathrm{ER}_x^{\mathrm{tot}}(y)}\mathrm{E}[N_x(y)|T_x(y)<1]\mathrm{E}[C_{x,1}(y)|T_x(y)<1] \\
& + \sum_{j=1}^{m-1} \frac{\mathrm{ER}_x^{j<T_x(y)\leq j+1}(y)}{\mathrm{ER}_x^{\mathrm{tot}}(y)}\mathrm{E}[N_x(y)|j<T_x(y)\leq j+1]\mathrm{E}[C_{x,1}(y)|j<T_x(y)\leq j+1] \\
& + \frac{\mathrm{ER}_x^{T_x(y)>m}(y)}{\mathrm{ER}_x^{\mathrm{tot}}(y)}\mathrm{E}[N_x(y)|T_x(y)>m]\mathrm{E}[C_{x,1}(y)|T_x(y)>m].
\end{aligned}$$

Notice that $L_x(y)$ disappears from the ratios of risk exposures and the weights assigned to the different categories only depend on the projected life table through one-year death probabilities $q_{x+j}(y+j)$ and j-year survival probabilities ${}_jp_x(y)$.

4.3 Numerical illustration

4.3.1 Data

Data has been collected and provided by the R&D department of one of the largest sickness fund operating in Belgium, the Christian Mutualities or ANMC. For confidentiality reasons, data have

been aggregated by attained age x, calendar year y and time to death $T_x(y)$, as explained earlier. Notice that this aggregation does not result in any loss of information with the proposed modeling.

The available database records hospital expenditures for 4.07 million Belgian inhabitants, representing about 41% of the whole population covered by the general regime. Note that one-day stays are excluded from the analysis, as well as psychiatric hospitalizations. Data contains information on patient's age, gender, date of death and inpatients hospital health expenditure. The response is the total expenditure (part covered by sickness fund together with out-of-pocket payments), broken down into a number of hospital stays and the corresponding amounts.

Explanatory variables included in the model are age x, $x \in \{70, 71, ..., 110\}$, and calendar year y, $y \in \{2002, ..., 2016\}$. The years $2002, ..., 2014$ are used as a training set. The observation period ends in 2016 so that we can allocate individuals to the different categories with respect to proximity to death. The model is fitted separately to male and female data, for each level of proximity to death. In that respect, we go beyond the classical decedents-versus-survivors setting that distinguishes between health care expenditures of individuals dying in a given period with those in the same age cohort who continue living. The model proposed in this chapter moves beyond the binary comparison of decedents and survivors by creating four categories reflecting end-of-life morbidity: those who die within the year, those who survive the next 12 to 24 months, those who survive over the next 24 to 36 months and those who survive at least 3 years. Moving from one category to the next helps to see how expenditures change as death approaches. Here, health status (or morbidity) is approximated by proximity to death. The last category is created to capture all individuals perceived as being in relatively good health (considering that health status cannot be predicted over a time horizon of 36 months). The proximity-to-death or remaining lifetime $T_x(y)$ is thus classified in 4 categories: $T_x(y) \leq 1$, $1 < T_x(y) \leq 2$, $2 < T_x(y) \leq 3$, and $T_x(y) > 3$. This corresponds to $m = 3$ in the general modeling strategy presented in the preceding section.

4.3.2 Descriptive statistics

Crude data are represented for the whole age range 0-100 on Figure 4.1. Considering the observed frequencies (displayed in the middle panels), we can see that average numbers of hospital admissions is higher around birth. It then decreases to reach a minimum, before increasing and then decreasing again at oldest ages. The marked peak around age 30 visible for females corresponds to hospital stays related to childbearing. Compared to mortality, the accident hump is present but not that visible because data are graphed on their original scale, and not on a logarithmic one.

Corresponding total yearly costs (displayed in the left panels), being influenced by their frequency component, also exhibit a similar age shape. Considering time trends, age-specific medical expenses also exhibit dynamic patterns because of the combined effect of longevity improvements and medical inflation. Also, the age-rationing effect is clear from the ultimate decrease visible on Figure 4.1. Its curving point coincides with a "normal life span" that can be inferred from life-expectancy at birth. Notice that the childbearing hump does not show up in the severities displayed in the right panels of Figure 4.1.

In this chapter, we model late-life health expenditures so that we restrict the age range to 70-100. Observed exposures are displayed in Figure 4.2, according to proximity to death. We can see there how risk exposures vary with age and proximity to death, separately for males and females. The age structure varies with proximity to death, as expected. These exposures enter the Poisson regression model, as volume measures.

Figures 4.3-4.5 represent frequencies, severities and total costs for the targeted age group 70-100, respectively, split according to proximity to death. Considering the range of values along the

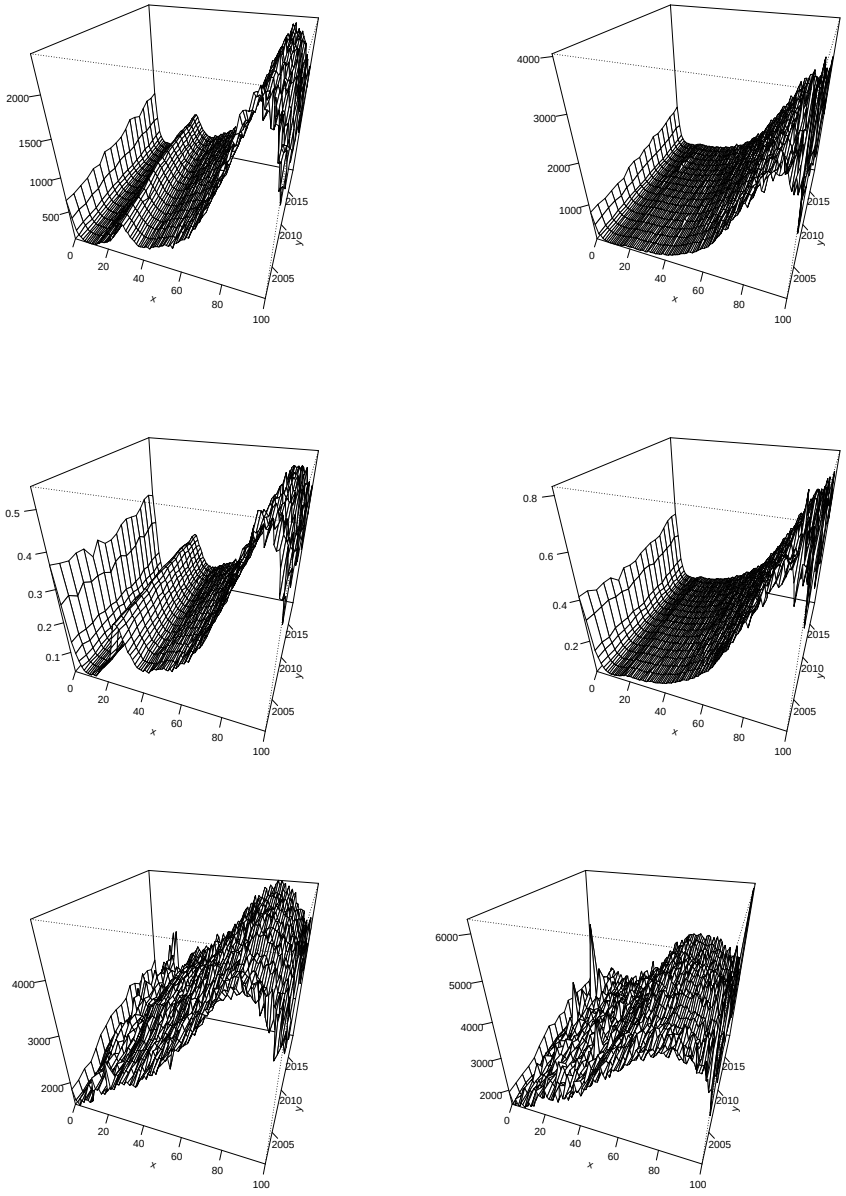

Figure 4.1: Observed average yearly costs $\overline{s}_x(y)$ (top panels), observed average yearly frequencies $\overline{n}_x(y)$ (middle panels), and observed average severities $\overline{c}_x(y)$ (bottom panels). Data for females appear on the left while data for males appear on the right.

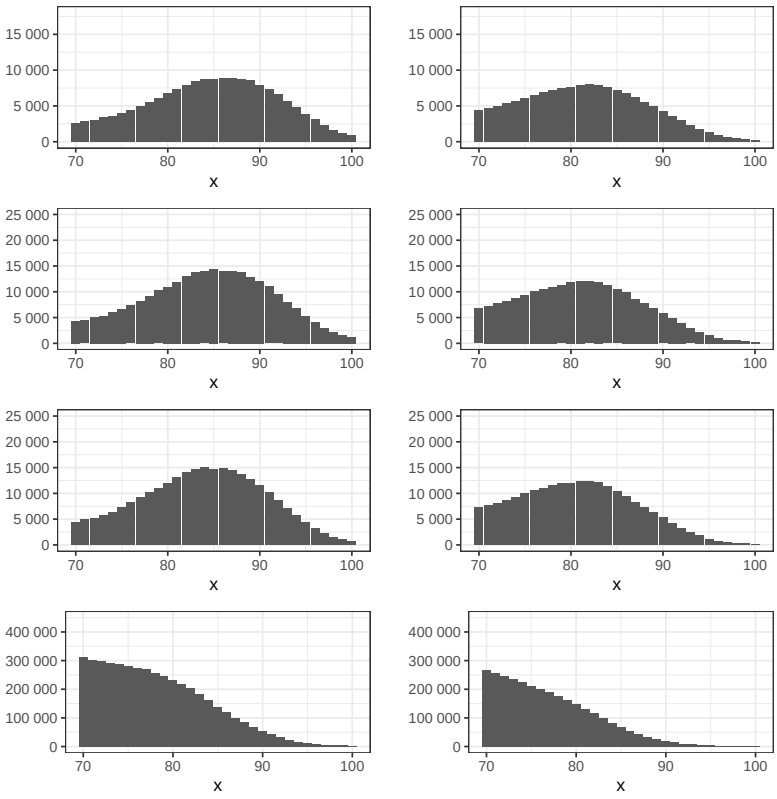

Figure 4.2: Risk exposures ER for females (left) and males (right). From top to bottom: $T_x(y) \leq 1$, $1 < T_x(y) \leq 2$, $2 < T_x(y) \leq 3$, and $T_x(y) > 3$.

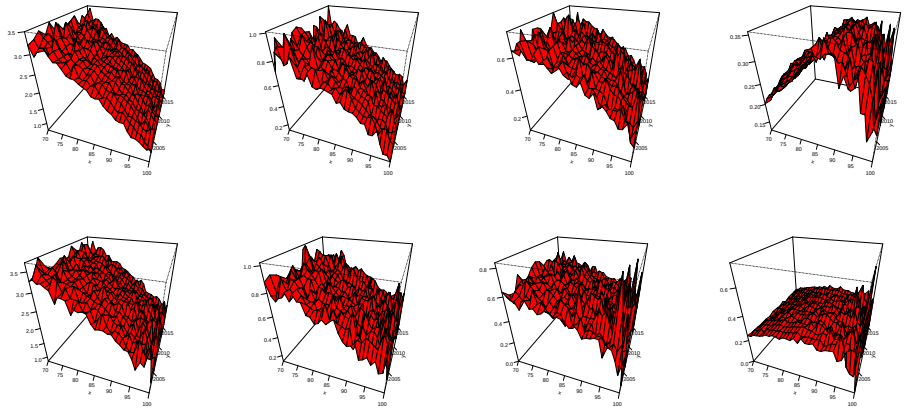

Figure 4.3: Observed average frequencies $\overline{n}_x(y)$ for females (top panels) and males (bottom panels). From left to right: $T_x(y) \leq 1$, $1 < T_x(y) \leq 2$, $2 < T_x(y) \leq 3$, and $T_x(y) > 3$.

z-axis, it is clear that decedents experience higher costs than survivors and the experience becomes more favorable with increasing time to death.

In terms of frequencies, we see on Figure 4.3 that the surfaces are profoundly modified when time to death increases. For individuals dying within the year (leftmost panels), average frequencies decrease with attained age, starting from a value around 2 at age 70. The decreasing shape is still visible for the two intermediate categories but for the individuals surviving the next three years (rightmost panels), the surface changes to an inverted U-shape with a peak around age 85 for males and 90 for females. The possible effect of calendar time is hard to assess on crude data.

The observed average severities displayed in Figure 4.4 are much more volatile compared to frequencies. It is nevertheless possible to see the decreasing trend in average costs per hospital stay at older ages for individuals who die within the year, in the leftmost panels. Both the average frequency and average severity components exhibit decreasing trends for decedents. The average cost for the final year of life thus decreases with attained age. The two intermediate categories do not reveal clear patterns. Considering individuals who survive the next three years (rightmost panels), data suggest an increasing trend with age but this is obscured by the huge volatility at oldest ages.

Figure 4.5 reveals that the shape of the yearly average total costs is similar to the shape of average frequencies displayed in Figure 4.3. There is no sign of time trend visible on Figure 4.5. As it will become clear in the next sections, this is due to the opposite trends in the frequency and severity components entering total costs.

4.3.3 Generalized Additive Modeling

Frequency model

Descriptive statistics show that age and time-until-death both have a clear impact on morbidity. Decedents notably experience a significantly higher number of hospitalizations, which then de-

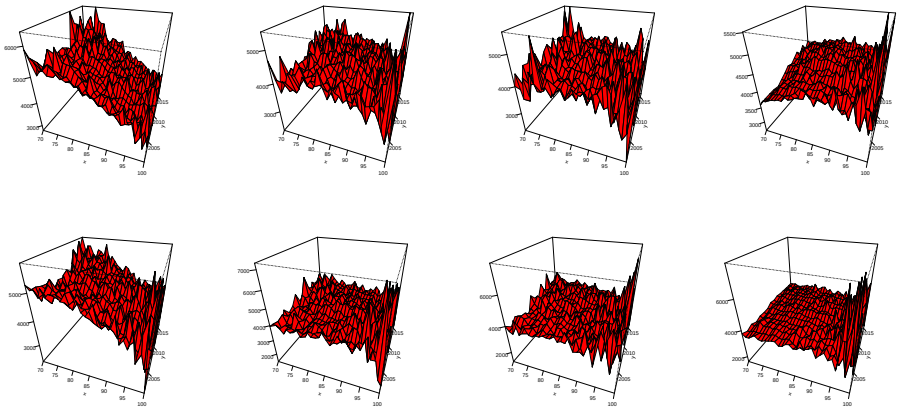

Figure 4.4: Observed average severities $\bar{c}_x(y)$ for females (top panels) and males (bottom panels). From left to right: $T_x(y) \leq 1$, $1 < T_x(y) \leq 2$, $2 < T_x(y) \leq 3$, and $T_x(y) > 3$.

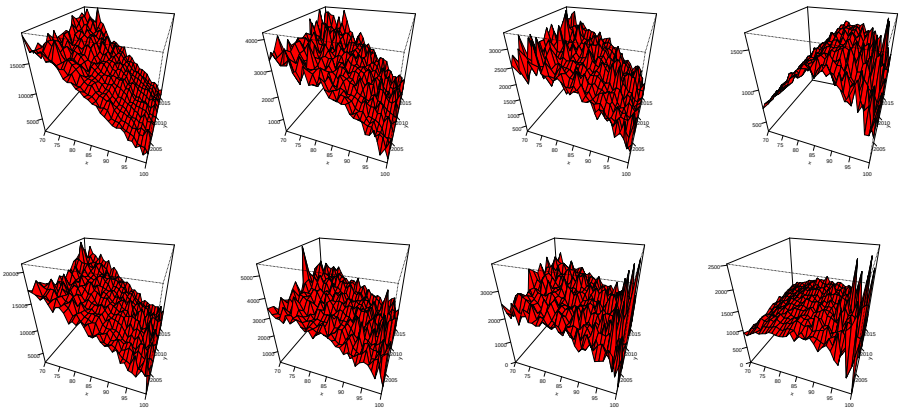

Figure 4.5: Observed average yearly total costs $\bar{s}_x(y)$ for females (top panels) and males (bottom panels). From left to right: $T_x(y) \leq 1$, $1 < T_x(y) \leq 2$, $2 < T_x(y) \leq 3$, and $T_x(y) > 3$.

Females					
Time to death	Degrees of freedom	p-value	R sq. adj.	deviance	AIC
$T_x(y) \leq 1$	20.03	$< 2 \cdot 10^{-16}$	99.6%	97.2%	3754.88
$1 < T_x(y) \leq 2$	15.07	$< 2 \cdot 10^{-16}$	98.5%	88.3%	3611.43
$2 < T_x(y) \leq 3$	12.27	$< 2 \cdot 10^{-16}$	98.6%	78.5%	3487.80
$T_x(y) > 3$	21.93	$< 2 \cdot 10^{-16}$	99.9%	97.9%	4100.98
Males					
Time to death	Degrees of freedom	p-value	R sq. adj.	deviance	AIC
$T_x(y) \leq 1$	14.11	$< 2 \cdot 10^{-16}$	99.6%	94.3%	3726.73
$1 < T_x(y) \leq 2$	18.87	$< 2 \cdot 10^{-16}$	98.8%	78.8%	3476.99
$2 < T_x(y) \leq 3$	11.92	$< 2 \cdot 10^{-16}$	98.9%	66.3%	3245.74
$T_x(y) > 3$	27.19	$< 2 \cdot 10^{-16}$	99.9%	87.7%	3916.24

Table 4.1: Summary of the nonlinear Poisson regression for the frequency component.

creases as time to death increases. To include age x and calendar year y into the frequency dynamics, separately in each category of proximity to death $j < T_x(y) \leq j+1$, we assume that the observed number of hospital admissions $N_x^{j<T_x(y)\leq j+1}(y)$ obeys the Poisson distribution with expected value equal to the observed exposure times the expected number $\lambda_x^{j<T_x(y)\leq j+1}(y)$ of hospital admissions per unit of exposure. Here, $\lambda_x^{j<T_x(y)\leq j+1}(y)$ is of the form $\ln \lambda_x^{j<T_x(y)\leq j+1}(y) = s(x,y)$ where $s(\cdot,\cdot)$ is a smooth function of x and y, to be estimated from the data. The smooth function $s(\cdot,\cdot)$ accounts for the effect of attained age x and calendar year y, on the logarithmic scale. This model is fitted to each gender and each proximity to death. The results are visible in Figure 4.6 for males and females. Table 4.1 summarizes the main information about the resulting fit in each category. All the calculations have been carried out with the help of the `mgcv` package of the statistical software R, contributed by Wood (2017).

The estimated functions $s(\cdot,\cdot)$ look similar for males and females. For those individuals who die in the next year, the fit is excellent as reflected in the adjusted R-squared and in the Deviance explained by the model. We can see on Figure 4.6 that the estimated surface is declining with attained age for decedents. This shows that the expected rate of admissions to hospital decreases as decedents age. Since the reduced exposures (because of death within the observation period) are accounted for, this may be attributed to age rationing. Actual-versus-Expected graphs are depicted in Figure 4.7 to assess the goodness of the fit. The fitted values appear along the x-axis whereas the observed ones correspond to the y-axis. The closer the pairs to the main diagonal, the better the fit. Proximity to the 45-degree line reveals the quality of the fit obtained with Poisson regression. The leftmost panels in Figure 4.7 reveal that the observed numbers of admissions are very well reconstituted by the regression model for decedents.

For those individuals who survive one year but die the following year or the ones that survive two years and die the third year, the fit remains very good. This is assessed by the high values of adjusted R-squared and by the Deviance explained by the model, respectively. Figure 4.7 shows that the observed counts are still very well explained by the model. The global trend is still declining with attained age, but an interaction captures the increase of the expected rate of admissions around 85 when longevity is improved.

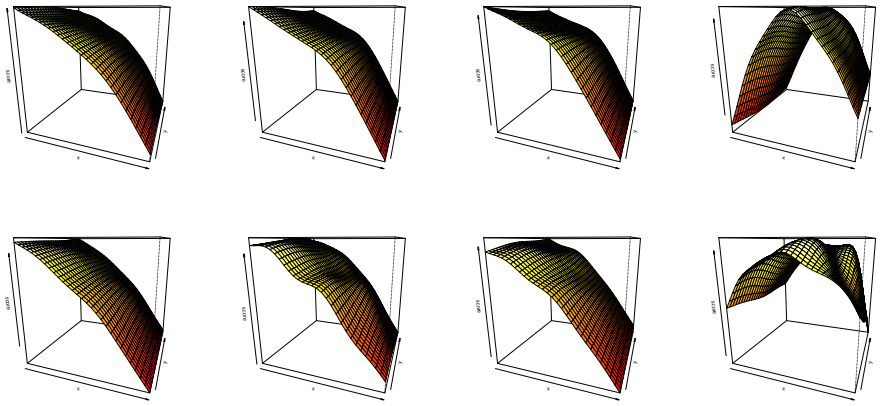

Figure 4.6: Estimated function s involved in the Poisson regression model for the numbers of hospital admissions, for females (top panels) and males (bottom panels). From left to right: $T_x(y) \leq 1$, $1 < T_x(y) \leq 2$, $2 < T_x(y) \leq 3$, and $T_x(y) > 3$.

Finally, considering the individuals who survive the next three years, the fit remains very good. Values of adjusted R-squared remain close to 1 and a large part of the Deviance is explained by the model. Figure 4.7 shows that the observed counts are very well explained by the model. Contrarily to those individuals who die earlier, we now obtain an inverse U-shape so that the trend is first increasing and then decreasing with attained age. Notice that these individuals can be considered to have a good health status.

Compared to the crude frequencies displayed in Figure 4.3, we see that the estimated expected number of hospital admissions obtained from the fitted surfaces displayed in Figure 4.6 well capture the structural pattern in age x and calendar year y, once random noise has been removed. The obtained surfaces appear to be very regular, except maybe for males surviving more than three years where a more complex dependency in x and y is visible. Figure 4.6 shows that the expected frequency sometimes declines over calendar time, depending on proximity to death and attained age. Overall, Figure 4.7 shows that fitted and actual values are in close agreement but that differences grow with attained age, because of the higher volatility resulting from reduced exposures at older ages.

Severity model

The probability density function of the Gamma distribution is right-skewed, with a sharp peak and a long tail to the right. These characteristics are often visible on empirical distributions of health expenses. This makes the Gamma distribution a natural candidate for modeling hospital expenses.

The Gamma regression model considered here falls in the GAM setting. Precisely, to include age x and calendar year y into the severity dynamics, separately in each category of proximity to death $j < T_x(y) \leq j+1$, we assume that the observed average cost per hospital admissions $\overline{C}_x^{j<T_x(y)\leq j+1}(y)$ obeys the Gamma distribution with mean value $\mu_x^{j<T_x(y)\leq j+1}(y)$, with $\ln \mu_x^{j<T_x(y)\leq j+1}(y) = s(x,y)$ for some function $s(\cdot,\cdot)$ to be estimated from the data. The number of hospital stays is included as

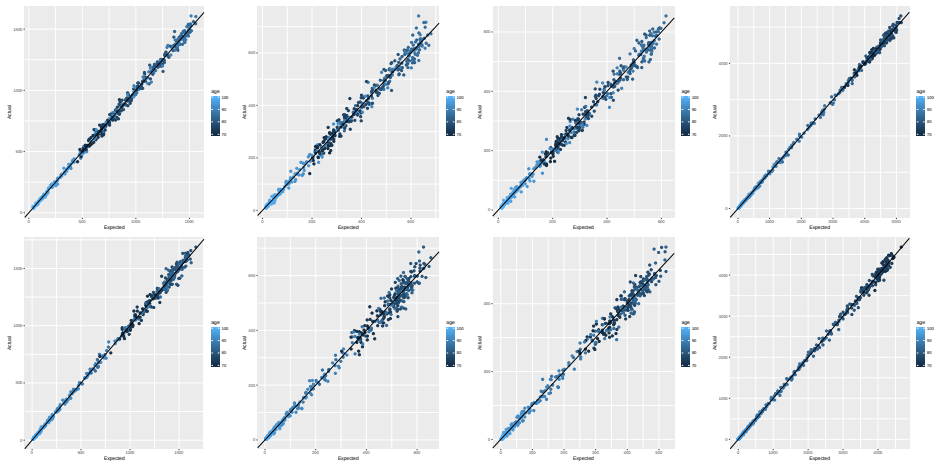

Figure 4.7: Actual versus expected values for the resulting Poisson regression fits for the frequency components. Females appear in the top panels, males in the bottom ones. From left to right: $T_x(y) \leq 1$, $1 < T_x(y) \leq 2$, $2 < T_x(y) \leq 3$, and $T_x(y) > 3$. Ages range from 70 (dark blue) to 100 (light blue).

a weight in the Gamma regression model. As for the Poisson regression for the numbers of hospital admissions, $s(\cdot,\cdot)$ is assumed to be a smooth function of attained age x and calendar year y.

The Gamma regression model is fitted to each gender and each proximity to death. The results obtained with the help of the `mgcv` package are visible in Figure 4.8 for males and females. Table 4.2 summarizes the main information about the resulting fit in each category. The effect of medial inflation is clearly visible on Figure 4.8 where we can see an increasing trend over the last observation periods. This is especially clear for individuals surviving for the next three years (rightmost panels). Considering the actual-versus-expected graphs displayed in Figure 4.9, the fit appears to be reasonable but of poorer quality compared to the frequency component, as reflected in the goodness-of-fit indicators listed in Table 2. This is due to the high volatility present in the data.

4.3.4 Comparison with an aggregate loss model

Compound Poisson sums are good candidates to model yearly total hospital expenses that are zero with positive probability, but continuously distributed otherwise. The Tweedie regression model correspond to responses of the form of compound Poisson sums with Gamma-distributed summands. It is thus in line with the Poisson specification for the number of hospital admissions and the Gamma specification for the cost per hospital stay used so far.

The Tweedie regression model also falls in the GAM setting. For each category of proximity to death $j < T_x(y) \leq j+1$, we assume that the observed total cost $S_x^{j<T_x(y)\leq j+1}(y)$ obeys the Tweedie distribution with mean value $\nu_x^{j<T_x(y)\leq j+1}(y)$, with $\ln \nu_x^{j<T_x(y)\leq j+1}(y) = s(x,y)$ for some function $s(\cdot,\cdot)$ to be estimated from the data. As before, $s(\cdot,\cdot)$ is assumed to be a smooth function of attained age x and calendar year y. Up to the observed exposures, $\nu_x^{j<T_x(y)\leq j+1}(y)$ corresponds to the prod-

Females					
Time to death	Degrees of freedom	p-value	R sq. adj.	deviance	AIC
$T_x(y) \leq 1$	22.47	$< 2 \cdot 10^{-16}$	86.5%	87.9%	4397438
$1 < T_x(y) \leq 2$	21.45	$< 2 \cdot 10^{-16}$	41.5%	44.8%	1880576
$2 < T_x(y) \leq 3$	17.84	$< 2 \cdot 10^{-16}$	35.8%	39.9%	1661929
$T_x(y) > 3$	20.47	$< 2 \cdot 10^{-16}$	85.2%	86.2%	12278051
Males					
Time to death	Degrees of freedom	p-value	R sq. adj.	deviance	AIC
$T_x(y) \leq 1$	20.2	$< 2 \cdot 10^{-16}$	70.2%	72.5%	4718747
$1 < T_x(y) \leq 2$	17.53	$< 2 \cdot 10^{-16}$	29.8%	33.4%	1831147
$2 < T_x(y) \leq 3$	10.78	$< 2 \cdot 10^{-16}$	32.8%	34.6%	1433268
$T_x(y) > 3$	17.74	$< 2 \cdot 10^{-16}$	72.5%	73.6%	8707526

Table 4.2: Summary of the nonlinear Gamma regression for the severity component.

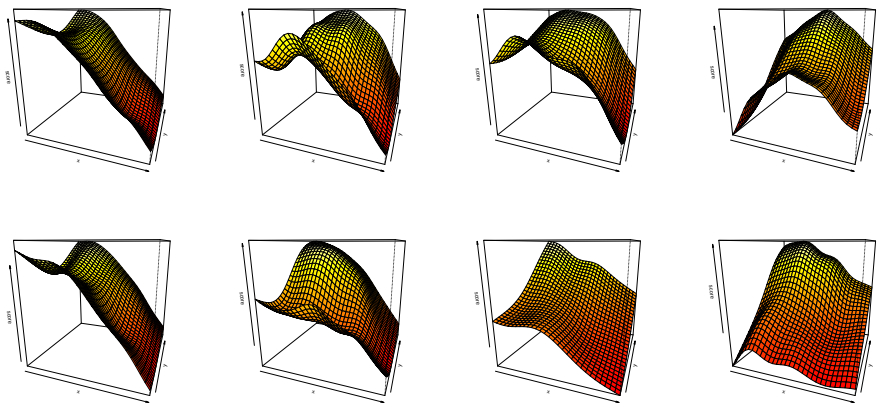

Figure 4.8: Estimated function s involved in the Gamma regression model for the average cost per hospital admissions, for females (top panels) and males (bottom panels). From left to right: $T_x(y) \leq 1$, $1 < T_x(y) \leq 2$, $2 < T_x(y) \leq 3$, and $T_x(y) > 3$.

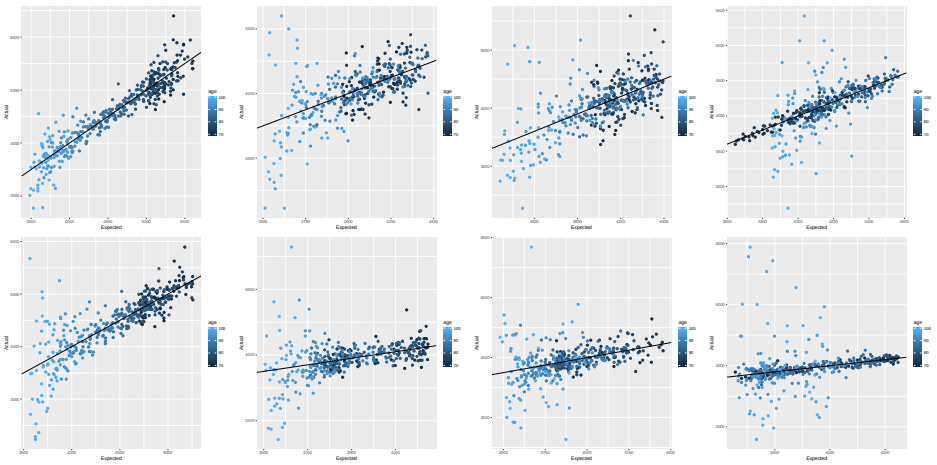

Figure 4.9: Actual versus expected values for the resulting Gamma regression fits for the severity components. Females appear in the top panels, males in the bottom ones. From left to right: $T_x(y) \leq 1$, $1 < T_x(y) \leq 2$, $2 < T_x(y) \leq 3$, and $T_x(y) > 3$. Ages range from 70 (dark blue) to 100 (light blue).

uct of $\lambda_x^{j<T_x(y)\leq j+1}(y)$ and $\mu_x^{j<T_x(y)\leq j+1}(y)$ introduced before so that the function $s(\cdot,\cdot)$ entering Tweedie regression appears to be the sum of the corresponding functions in Poisson and Gamma regression models.

Tweedie regression model is fitted to each gender and each proximity to death. The results are visible in Figure 4.10 for males and females. All the calculations have been carried out with the help of the `mgcv` package, as before. Table 4.3 summarizes the main information about the resulting fit in each category. Considering the actual-versus-expected graphs displayed in Figure 4.9, the fit is generally good, except for old people, see Figure 4.11.

4.3.5 Out-of-sample analysis

Figures 4.12 and 4.13 compare expected total costs with observed ones for the last years of observation comprised in the database that have not been used for model training, i.e. for 2015-2016 and the years 2017-2019 for which individuals cannot be classified into categories defined according to the proximity to death. We can see that there are some moderate departures between observations and fitted values obtained from the Tweedie aggregated model and the frequency-severity approach at older ages. Overall, the two approaches produce results in close agreement. We will see below that the differences between the two approaches materialize in the projections.

4.3.6 Projections

Let us now project expected hospital costs beyond the end of the observation period, thus for calendar years posterior to 2016, with the help of the Poisson and Gamma regression models combined together, on the one hand, and of the Tweedie regression model, on the other hand. Plugging future

Females					
Time to death	Degrees of freedom	p-value	R sq. adj.	deviance	AIC
$T_x(y) \leq 1$	28.98	$< 2 \cdot 10^{-16}$	97.2%	97.6%	3115355
$1 < T_x(y) \leq 2$	28.97	$< 2 \cdot 10^{-16}$	85.6%	87.6%	4163553
$2 < T_x(y) \leq 3$	28.97	$< 2 \cdot 10^{-16}$	73.1%	76.9%	4024641
$T_x(y) > 3$	29	$< 2 \cdot 10^{-16}$	96.7%	97.3%	46458467
Males					
Time to death	Degrees of freedom	p-value	R sq. adj.	deviance	AIC
$T_x(y) \leq 1$	28.96	$< 2 \cdot 10^{-16}$	93.4%	94.1%	2920664
$1 < T_x(y) \leq 2$	28.96	$< 2 \cdot 10^{-16}$	71.4%	74.5%	3692354
$2 < T_x(y) \leq 3$	28.95	$< 2 \cdot 10^{-16}$	61.4%	64.7%	3363281
$T_x(y) > 3$	29	$< 2 \cdot 10^{-16}$	79.4%	81.3%	32930849

Table 4.3: Summary of the nonlinear Tweedie regression for total costs.

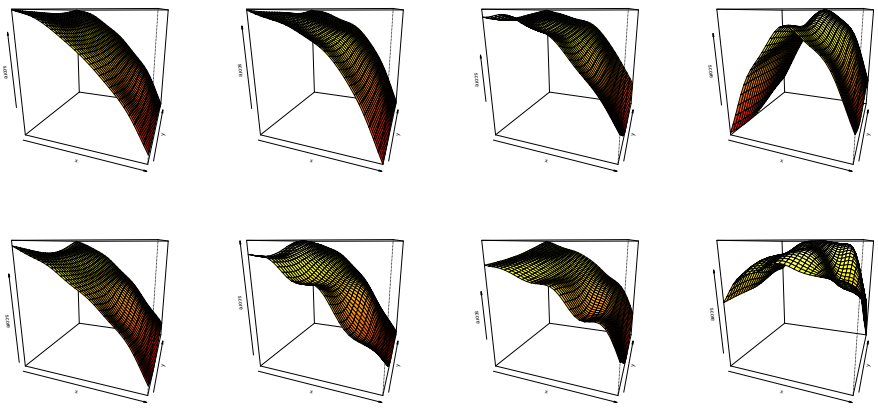

Figure 4.10: Estimated function s involved in the Tweedie regression model for total costs, for females (top panels) and males (bottom panels). From left to right: $T_x(y) \leq 1$, $1 < T_x(y) \leq 2$, $2 < T_x(y) \leq 3$, and $T_x(y) > 3$.

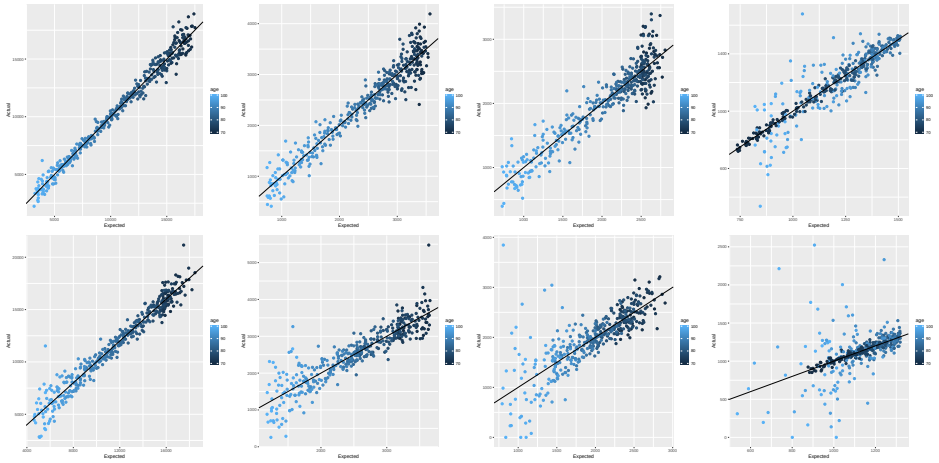

Figure 4.11: Actual versus expected values for the resulting Tweedie regression fits for total costs. Females appear in the top panels, males in the bottom ones. From left to right: $T_x(y) \leq 1$, $1 < T_x(y) \leq 2$, $2 < T_x(y) \leq 3$, and $T_x(y) > 3$. Ages range from 70 (dark blue) to 100 (light blue).

calendar years in the nonlinear regression model provides the analyst with the forecast of hospital expenditures over the next years.

Results for the different proximity categories are combined with the help of expected exposures for future years, as explained in Section 2. Projections are performed over 2020-2030 in Figure 4.14. Both approaches, based on Tweedie regression and frequency-severity decomposition, forecast decreasing expected total costs over all ages. The difference between the two approaches materializes in the frequency and severity projections, as different trends clearly emerge.

In order to make frequency projections, results for the different proximity categories are combined with the help of expected exposures for future years, as mentioned above. Frequency projections are performed over 2020-2030 in Figure 4.15. Both male and female predictions show a decrease in expected frequency over all ages, also postulated by e.g. Fries (1983) or Miller (2001). At the age 85 for example, the relative expected decrease between 2020 and 2030 is 7% for males and 13% for females. This decrease is also corroborated by practitioners and can be explained in different ways. It confirms the 'red herring' hypothesis, which postulates that high end-of-life costs are being postponed with the increasing longevity, inducing a downward shifting in old age hospitalization costs with time. This is confirmed by the observed global increase in healthy life expectancy or disability free life expectancy. In Belgium this increase is estimated between 2004 and 2018 at 2,7 years for males and 1,4 years for females, see Sciensano (2019). Another explanation is related to the substitution of classical hospitalization stays by one-day stays, which are not accounted for in the data. Reasons for this substitution comprise medical technical progress, better health and improved check-ups. The recent development of autonomy and dependency ambulatory care can also explain the reduction in classical hospitalization needs.

In order to make severity projections, results for the different proximity categories are combined

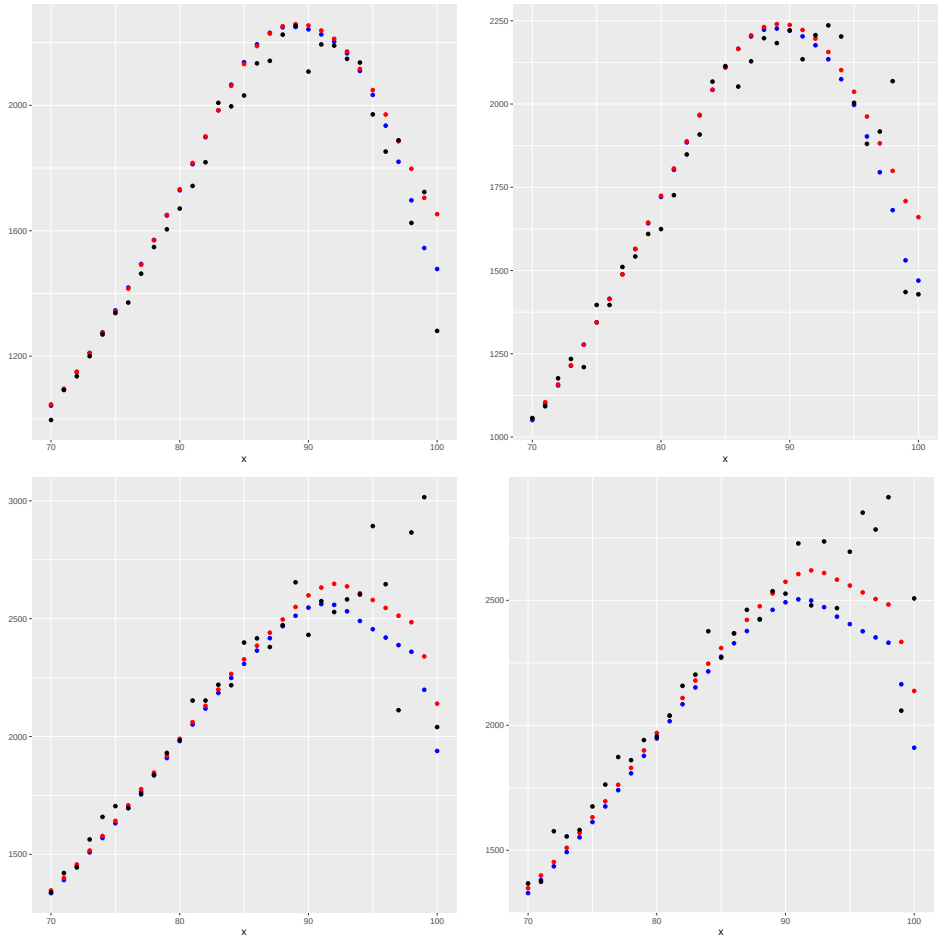

Figure 4.12: Observed total costs (black dots) versus fitted values for females (top) and males (bottom), calendar year 2015 (left panels) and 2016 (right panels). Fitted values obtained with frequency-severity decomposition appear printed in red whereas those obtained from Tweedie regression appear printed in blue.

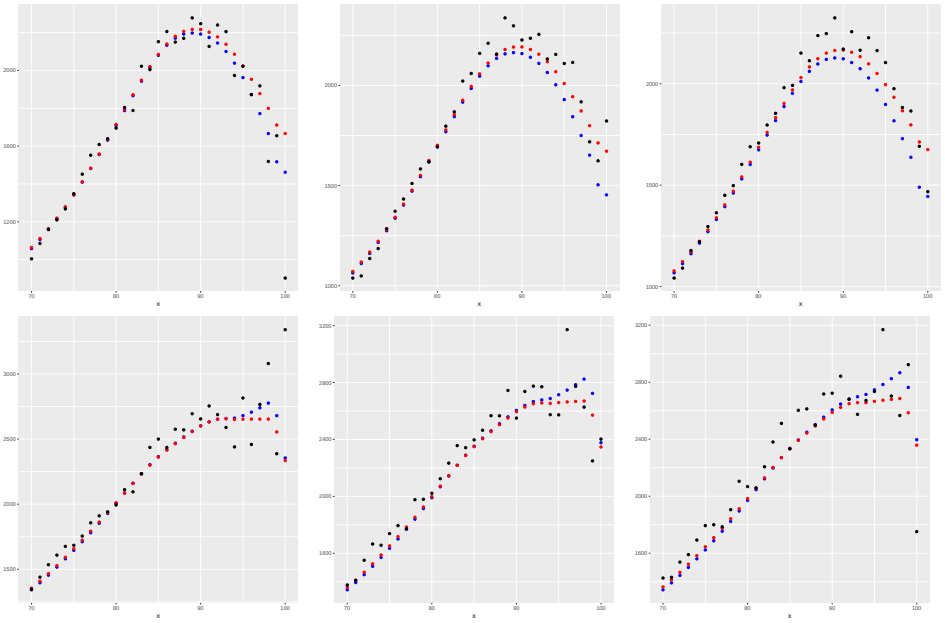

Figure 4.13: Observed total costs (black dots) versus fitted values for females (top) and males (down), calendar year 2017 (left panels), 2018 (middle panels) and 2019 (right panels). Fitted values obtained with frequency-severity decomposition appear printed in red whereas those obtained from Tweedie regression appear printed in blue.

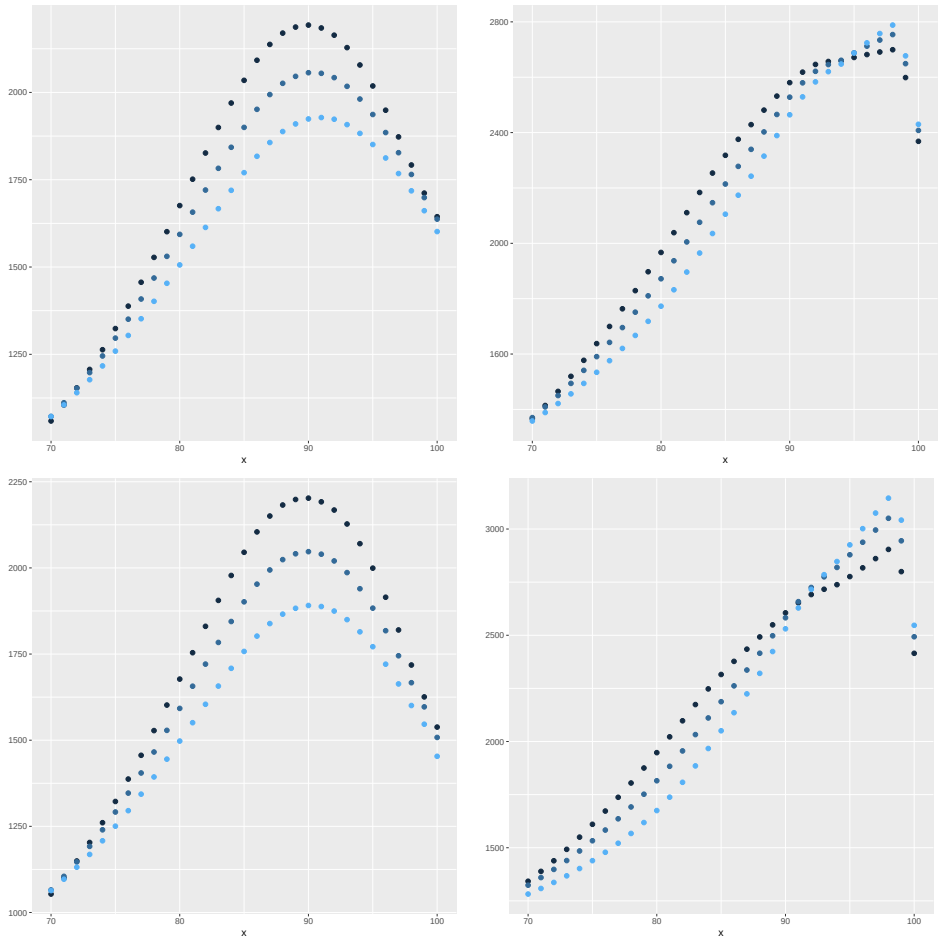

Figure 4.14: Forecast of expected total cost $S_x(y)$ for female (left) and male (right), year 2020, 2025 and 2030 (from dark to light blue). Frequency-severity decomposition (top panels) and Tweedie total cost (bottom panels).

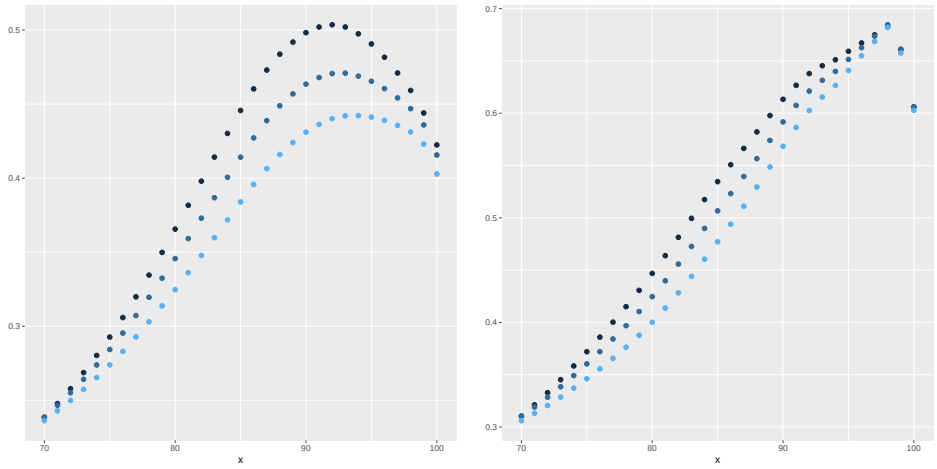

Figure 4.15: Forecast of expected frequency $\lambda_x(y)$ for female (left) and male (right), year 2020, 2025, 2030 (from dark to light blue).

using:

$$E[\overline{C}_x(y)] = \frac{\sum_{j=0}^{2} \mu_x^{j<T_x(y)\leq j+1}(y) \cdot \lambda_x^{j<T_x(y)\leq j+1}(y) + \mu_x^{T_x(y)>3}(y) \cdot \lambda_x^{T_x(y)>3}(y)}{\sum_{j=0}^{2} \lambda_x^{j<T_x(y)\leq j+1}(y) + \lambda_x^{T_x(y)>3}(y)}$$

with $\overline{C}_x(y)$ the expected average cost per hospital stay at age x. Severity projections are performed over 2020-2030 in Figure 4.16. Female forecasts do not show any clear trend while male forecasts show a positive trend. This mainly relates to the inflation effect, which is almost null for females.

In total health care expenditures projections, a global decrease of the costs is expected. Two competing effects can be separately captured in the frequency-severity decomposition approach. A decreasing and dominating trend is induced by the frequency component, while a positive or stable effect is induced by the severity component.

4.4 Conclusion

Breaking total inpatient hospital costs into a frequency and a severity component greatly helps to understand the underlying dynamics across age and time. Incorporating age, calendar time and proximity-to-death represents an important advance over simple age-based models, especially for the elderly who are subject to high death rates. The inclusion of proximity to death also allows the analyst to include longevity projection in the forecast of future costs, by means of projected life tables.

The main findings of this chapter are as follows: expected annual hospitalization cost is experimenting a global decreasing trend, alongside an increasing longevity. The main

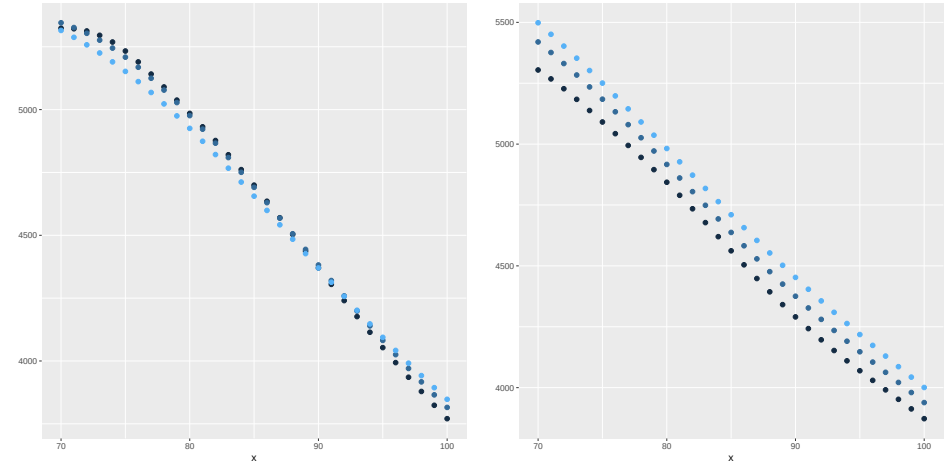

Figure 4.16: Forecast of expected severity $\mu_x(y)$ for female (left) and male (right), year 2020, 2025, 2030 (from dark to light blue).

driver is the frequency component. The severity upward or stable trend, i.e. inflation effect, is not strong enough to counterbalance the reduced morbidity. The particular role of proximity-to-death on hospitalization costs postulated by e.g. Miller (2001) has been confirmed and the special case of the decedent category, i.e. dying within the year, has been highlighted. The proposed frequency-severity decomposition model was compared to the aggregate Tweedie model and it has the clear advantage to disentangle individual competing effects, allowing better predictions.

Even if the model has been fitted here on aggregate data, for confidentiality reasons, it could theoretically be calibrated on individual data as well so that the time to death could also be treated as a continuous feature and not as a categorical one.

The severity could be reformulated as the product between the number of inpatient days and the daily inpatient cost. Therefore the severity random variable $C_{x,k}(y)$ for the kth stay of a person aged x and time y could be further decomposed as $C_{x,k}(y) = d_{x,k}(y) \times c_{x,k}(y)$, where $d_{x,k}(y)$ is the length of stay for the kth hospital claim and $c_{x,k}(y)$ the daily hospital cost. A specific trend as regards hospital stays duration could come out and this would help better capture severity trend. There is indeed an actual trend to reduce the number of inpatient days (for a same pathology), which impacts the severity component. As previously stated classical hospital stays are being increasingly substituted by one day stays, which relates mainly to medical progress.

To end with, let us mention that the future observed 2020 hospitalization data will naturally be impacted by the global COVID pandemic. The proposed models and their predictions did not consider any external shock, like a catastrophe or pandemic. It is expected that total hospitalization claims will globally decrease in 2020, see e.g. IMA (2020) and Solidaris (2020), but a recoup effect is expected in 2021, so that medium-term

predictions 2025-2030 should remain meaningful in case the COVID episode only has transitory effects.

References

- Agence Intermutualiste (IMA). (2020) Impact de la crise du coronavirus sur le nombre d'hospitalisations. https://aim-ima.be/Impact-de-la-crise-du-coronavirus.

- Breyer, F. & Felder, S. (2006). Life expectancy and health care expenditures: A new calculation for Germany using the costs of dying. Health Policy 25, 178-186.

- Christiansen, M., Denuit, M., Lucas, N. & Schmidt, J.-Ph. (2018). Projection models for health expenses. Annals of Actuarial Science 12, 185-203.

- Direction générale Statistique Statistics Belgium, Bureau Fédéral du Plan (2016). Perspectives démographiques 2015-2060 Population, ménages et quotients de mortalité prospectifs. Available from http://www.plan.be/

- Felder, S., Werblow, A. & Zweifel, P. (2010). Do red herrings swim in circles? Controlling for the endogeneity of time to death. Journal of Health Economics 29, 205-212.

- Frees, E., Gao, J. & Rosenberg, M. (2011). Predicting the frequency and amount of health care expenditures. North American Actuarial Journal 15, 377-392.

- Fries, J. (1983). The compression of morbidity. The Milbank Memorial Fund Quarterly. Health and Society, 397-419.

- Miller, T. (2001). Increasing longevity and Medicare expenditures. Demography 38, 215-226.

- Payne, G., Laporte, A., Deber, R. & Coyte, P. (2007). Counting backward to health care's future: Using time-to-death modeling to identify changes in end-of-life morbidity and the impact of aging on health care expenditures. The Milbank Quarterly 85, 213-257.

- Schokkaert, E. & Van de Voorde, C. (2003). Belgium: Risk adjustment and financial responsibility in a centralised system. Health Policy 65, 5-19.

- Sciensano (2019). Espérance de Vie et Qualité de Vie. https://www.belgiqueenbonnesante.be /fr/etat-de-sante/esperance-de-vie-et-qualite-de-vie/esperance-de-vie-en-bonne-sante

- Solidaris (2020). COVID-19 : quel impact sur la médecine générale? Available from http://www.solidaris.be

- Stearns, S., & Norton, E. (2004). Time to include time to death? The future of health care expenditure predictions. Health Economics 13, 315-327.

- Werblow, A., Felder, S. & P. Zweifel (2007). Population ageing and health care expenditure: A school of "Red Herrings"? Health Economics 16, 1109-1126.

- Wood, S. (2017). Generalized Additive Models: An Introduction with R. Second edition. Chapman and Hall/CRC.

- Zweifel, P., Felder, S. & Meiers, M. (1999). Ageing of population and health care expenditure: A red herring? Health Economics 8, 485-496.

- Zweifel, P., Felder, S. & Werblow, A. (2004). Population ageing and health care expenditure: New evidence on the "red herring". The Geneva Papers on Risk and Insurance Issues and Practice 29, 652-666.

Chapter 5

Part II: conclusion

Systematic risks are comprised in the classical hospitalization cover and inflation for instance is the main driver in Chapter 2. In Chapter 3 aggregate health related costs are modeled in a stochastic way similarly to mortality. The time factor in the proposed Lee-Carter approach would relate to the age-independent annual health claim trend. Therefore the drift θ would closely relate to the inflation denoted $(1+f^{(k)})$ in Chapter 2. It is linked to the morbidity evolution and comprises notably the inflationary or severity evolution but also a demographic factor.

In Chapter 2 the annual medical inflation f_k is supposed to be age-independent and it triggers an annual increase of expected annual cost such that $b_x^{(k)} = b_x^{(k-1)}(1+f_k)$, with $b_x^{(k)}$ the expected annual medical claim at age x based on hypotheses computed at time k. The Lee-Carter drift θ is also age-independent but the medical cost increase is supposed to vary with age, thanks to the β_x parameter. The parameter θ is estimated by the model and is based on linear smoothing. The medical inflation f_k is computed each year and corresponds to the Belgian national medical inflation indice, computed as in SPF (2019)(see also Devolder et al. (2008)). It is therefore an ex-post quantity subject to random evolution. The θ drift parameter in Chapter 3 is estimated based on male German private insurance data and therefore is not fully in line with the Belgian medical inflation $f^{(k)}$ in Chapter 2.

The results in Chapter 3 suggest a rather stable per-age cost trend (cf. Figure 3.13), corroborating Chapter 4, which suggests a stable or even decreasing trend. The results in both Chapters should be compared with adequate attention as they relate to very different datasets, different population, different genders (i.e. Chapter 3 only looks into male while Chapter 4 looks at both genders), different ages and different years. Indeed Chapter 3 relates to the German private insurance market and encompasses ages from 20 to 80 with an available history 1995-2011 while in Chapter 4 the data originates from the Belgian public system and targets very old people, i.e. 70-110 with an available history 2002-2016. Both models are supposed to perform well on a very short-term horizon due to the

limited available history, but are not meant to be powerful long-term prediction tools.

The results could actually be reconciliated. Figures 3.13 and 3.14 in Chapter 3 show a general stable health claims evolution like the model proposed in Chapter 4. If we look into male results and compare Lee-Carter parameters for ages over 70 (positive β_x s and increasing predicted κ_t s) with the predicted results of the Chapter 4 regression model (see Figure 4.14 on page 76), the trend is actually slightly positive in both approaches.

References

1. SPF, L'indice médical. https://economie.fgov.be/fr/themes/services-financiers/assurances/

2. Devolder, P., Denuit, M., Maréchal, X., Yerna, B-L, Closon, J-P, Léonard C., et al. (2008) Construction d'un index médical pour les contrats privés d'assurance maladie. Health Services Research (HSR). Bruxelles: Centre fédéral d'expertise des soins de santé. KCE reports 96B

Part III

Long-term care insurance

Chapter 6

Insurance approach

The present chapter is based on the following edited book chapter:

- Denuit, M., Lucas, N., Pitacco, E. (2019). Pricing and Reserving in LTC Insurance. In Dupourqué, E., Planchet, F., Sator, N. (eds) 'Actuarial Aspects of Long Term Care', pp 129-158. Springer Actuarial.

6.1 Introduction

This chapter aims to present the actuarial calculation techniques for pricing and reserving in LTC insurance products providing predefined benefits, thus disregarding in particular cost reimbursement benefits. The multistate structure is consistent with such predefined benefits but also allows to quantify the time LTC is needed, i.e. the duration of payment of insurance benefits of any type. The model used here is a hierarchical 3-state model. We do not consider temporary loss of autonomy but assume that reactivation is not possible: there is thus a single, irreversible state of dependence. The absence of recovery is justified because only people with severe disability are eligible for LTC benefits, in general. The semi-Markov framework in the LTC state relies on two variables: age and continuance or occupation time. One should remind that there is an important heterogeneity in mortality among LTC beneficiaries, according to major types of pathologies inducing the LTC need, but it will not be considered here.

The chapter is organized as follows. Section 5.2 describes the multistate modeling for LTC insurance policies. The quantities entering actuarial calculations (transition probabilities and intensities) are defined in Sections 5.3 and 5.4. The actuarial equivalence principle is applied to LTC insurance pricing in Section 5.5. Sections 5.6 to 5.9 present analytical expressions for premiums related to some specific LTC insurance products. Combined products are also considered. Section 5.10 discusses the reserving process and provides the reader with some analytical expressions. The final Section 5.11 concludes the chapter.

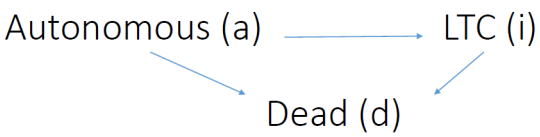

Figure 6.1: Multistate model for the LTC insurance cover.

6.2 Multistate modeling

Multistate models provide a convenient representation for life and health insurance contracts, including LTC. Each state represents a particular status for the policyholder. The benefits comprised in the contract are associated to sojourns in, or transitions between states. See, e.g., Chapter 8 in Dickson et al. (2013) or Pitacco (2014) for an introduction.

In the remainder of this chapter, we consider a three-state model, and the following notation is adopted. Henceforth, x denotes policyholder's age at policy issue. We assume that there is an ultimate age $\omega \leq \infty$ and we denote as

$$\omega_x = \omega - x$$

the maximal time until death for an individual aged x. Time t measures time since policy issue and thus corresponds to contract seniority.

Policyholder's history is described by the stochastic process $\{X_t,\ t \geq 0\}$ where X_t gives the state occupied by the individual at time t, with $X_t \in \{\text{a,i,d}\}$ as shown in Figure 6.1, where state a stands for "autonomous" or "active", state i stands for "invalid" or "disabled", and state d stands for "dead". The LTC state where benefits are paid thus corresponds to i. Henceforth, only transitions a$\rightarrow$i, a$\rightarrow$d and i$\rightarrow$d are allowed so that the loss of autonomy is assumed to be permanent (no recovery possible).

This non-reversibility greatly simplifies the calculations (as the 3-state process is hierarchical and trajectories can easily be described in terms of just a few random variables) and appears to be reasonable at older ages (at which the LTC need becomes stronger). In case recoveries are possible, calculations can be carried on using the so-called Waters algorithm based on time discretization. We refer the reader to Waters (1990) for further details about the algorithm.

The time spent in the LTC state i influences future transitions. This is why we introduce the random variable Z_t defined as the time spent in the state occupied at time t, i.e.

$$Z_t = \max\{z \leq t | X_t = X_{t-h} \text{ for all } 0 \leq h \leq z\}.$$

It is assumed that only the current state X_t and the time Z_t spent in the current state influence future transitions so that $\{X_t, t \geq 0\}$ is a semi-Markov process, i.e. $\{(X_t, Z_t), t \geq 0\}$ is a Markov process. Notice that only the LTC state i requires the semi-Markov assumption,

i.e. probabilities of future transitions from that state also depend on the occupation time Z_t.

6.3 Transition probabilities

We consider a policyholder who is autonomous and aged x at policy issue, i.e. we work conditionally on $X_0 = \mathrm{a}$. The following transition probabilities are needed in the three-state LTC model:

$$\begin{aligned}
{}_up_{x+t}^{\mathrm{ai}} &= \mathrm{P}[X_{t+u} = \mathrm{i} | X_t = \mathrm{a}] \\
&= \text{probability for an individual in state a at time } t \text{ of being in state i} \\
&\quad \text{at time } t+u \\
{}_up_{x+t}^{\mathrm{ad}} &= \mathrm{P}[X_{t+u} = \mathrm{d} | X_t = \mathrm{a}] \\
&= \text{probability for an individual in state a at time } t \text{ of being in state d} \\
&\quad \text{at time } t+u \\
{}_up_{x+t;z}^{\mathrm{id}} &= \mathrm{P}[X_{t+u} = \mathrm{d} | X_t = \mathrm{i}, Z_t = z] \\
&= \text{probability for an individual in state i at time } t \text{ since time } t-z \\
&\quad \text{of being in state d at time } t+u \\
{}_up_{x+t}^{\mathrm{aa}} &= \mathrm{P}[X_{t+u} = \mathrm{a} | X_t = \mathrm{a}] \\
&= \text{probability for an individual in state a at time } t \text{ of being in state a} \\
&\quad \text{at time } t+u \\
{}_up_{x+t;z}^{\mathrm{ii}} &= \mathrm{P}[X_{t+u} = \mathrm{i} | X_t = \mathrm{i}, Z_t = z] \\
&= \text{probability for an individual in state i at time } t \text{ since time } t-z \\
&\quad \text{of being in state i at time } t+u.
\end{aligned}$$

The Semi-Markov assumption ensures that these transition probabilities entirely describe the distribution of the stochastic process $\{X_t, t \geq 0\}$ giving policyholder's individual experience.

By assumption, recovery is not possible. Hence, transition probabilities ${}_up_{x+t}^{\mathrm{aa}}$ and ${}_up_{x+t;z}^{\mathrm{ii}}$ are in reality sojourn probabilities, i.e.

$$\begin{aligned}
{}_up_{x+t}^{\mathrm{aa}} &= \mathrm{P}[X_{t+h} = \mathrm{a} \text{ for all } 0 < h \leq u | X_t = \mathrm{a}] \\
{}_up_{x+t;z}^{\mathrm{ii}} &= \mathrm{P}[X_{t+h} = \mathrm{i} \text{ for all } 0 < h \leq u | X_t = \mathrm{i}, Z_t = z].
\end{aligned}$$

6.4 Transition intensities

Transition intensities quantify the instantaneous risk of making a given transition, depending on the state currently occupied. They extend the force of mortality at the heart of

life insurance mathematics to more general multistate models describing health insurance products, including LTC ones.

From the above transition probabilities, the transition intensities are derived by the following limits:

$$\begin{aligned}
\mu^{\mathrm{ai}}_{x+t} &= \lim_{h\searrow 0} \frac{{}_hp^{\mathrm{ai}}_{x+t}}{h} \\
\mu^{\mathrm{ad}}_{x+t} &= \lim_{h\searrow 0} \frac{{}_hp^{\mathrm{ad}}_{x+t}}{h} \\
\mu^{\mathrm{id}}_{x+t;z} &= \lim_{h\searrow 0} \frac{{}_hp^{\mathrm{id}}_{x+t;z}}{h}, \qquad z<t.
\end{aligned}$$

As state a remains Markovian, the transition intensities from that state do not depend on the time spent there, but only on attained age $x+t$. On the contrary, there is an effect of the duration of stay in state i so that transition intensities from i depend on both attained age $x+t$ and time z spent in the LTC state.

Transition rates are often assumed to be piecewise constant. This assumption greatly eases the actuarial calculations. There are essentially two approaches to make the Semi-Markov transition intensities $(y,z)\mapsto \mu^{\mathrm{id}}_{y;z}$ piecewise constant:

- either transitions intensities vary at integer ages and sojourn duration in the LTC state, i.e. for every integer y and z,

  $$\mu^{\mathrm{id}}_{y+\xi;z+s} = \mu^{\mathrm{id}}_{y;z} \text{ for all } 0\le \xi < 1 \text{ and } 0\le s<1.$$

  Of course, finer grid can be used (this is often useful for the LTC state, where death rates vary rapidly during the first year after the loss of autonomy).

- or specific, piecewise constant transition rates apply according to the age at entry in the LTC state, i.e.

  $$\mu^{\mathrm{id}}_{y+\xi;z} = \widetilde{\mu}(y+\lfloor \xi - z\rfloor, \lfloor z\rfloor)$$

  for some given function $\widetilde{\mu}$ defined on $\mathbb{N}^2$, where $\lfloor\cdot\rfloor$ denotes the integer part. The arguments of $\widetilde{\mu}(\cdot,\cdot)$ are age at loss of autonomy and time spent in the LTC state, respectively.

The second approximation is very convenient for the computations. Rates are displayed in a matrix: the age (last birthday) at loss of autonomy is the first dimension while the time since occurrence is the second dimension (the sum of these two values giving the attained age). We retain the second approximation in this chapter.

Transition intensities are displayed in Figures 6.2-6.4. They correspond to values in line with observations made on the French LTC market. We see that μ^{ai}_y and μ^{ad}_y exponentially increase with age y. Considering the death rate in the LTC state, notice that age on the graph corresponds to the age at entry in state i so that individuals are subject to

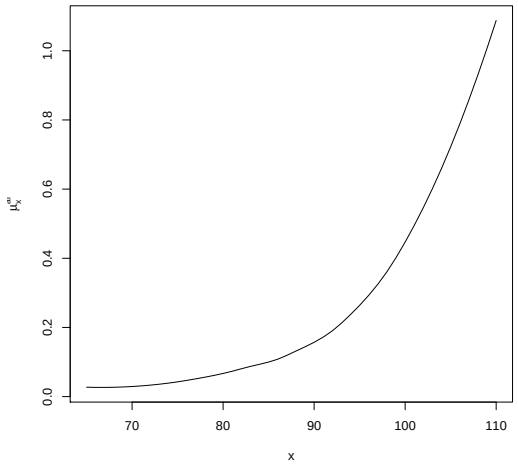

Figure 6.2: Transition rate $y \mapsto \mu_y^{\mathrm{ai}}$.

death rates $\mu_{y+z;z}^{\mathrm{id}} = \widetilde{\mu}(y,z)$ if they lost autonomy at age y. We can see on Figure 6.4 that mortality is particularly high just after the loss of autonomy (i.e. for small values of z) and then decreases once the individual survived the first years spent in the LTC state.

We also define the exit rate from state a as

$$\mu_y^{\mathrm{a}\bullet} = \mu_y^{\mathrm{ai}} + \mu_y^{\mathrm{ad}}.$$

Clearly, the exit rate is also piecewise constant when μ_y^{ai} and μ_y^{ad} both exhibit this feature. The set of transition intensities form the analogous to the life table in life insurance, allowing the actuary to assign a probability to every event in relation to the LTC cover.

6.5 Equivalence principle

This principle used to compute life insurance premiums extends to all health insurance products. It states that at policy issue, the expected present value of the benefits paid to the policyholder is equal to the expected present value of the premiums paid to the insurer. The discount factor $v(s,t)$ is the present value at time s of a unit payment made at time t, $s < t$, with $v(s,s) = 1$. In the numerical illustrations, we assume that the technical interest rate is constant over time, i.e.

$$v(s,t) = \exp\big(-\delta(t-s)\big) \tag{6.5.1}$$

for some $\delta > 0$.

The benefits comprised in LTC policies are as follows:

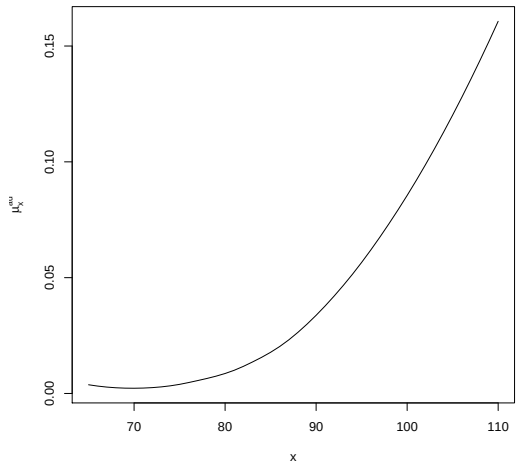

Figure 6.3: Transition rate $y \mapsto \mu_y^{\mathrm{ad}}$.

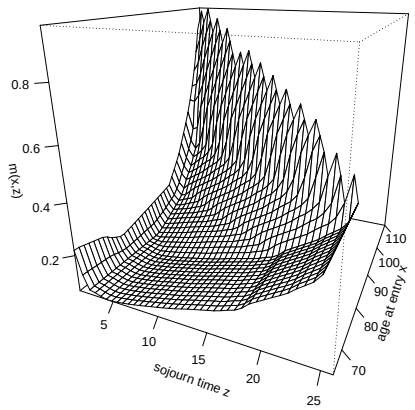

Figure 6.4: Function $(y,z) \mapsto \widetilde{\mu}(y,z)$ defining death rates in the LTC state i.

b_{i} rate of time-continuous benefits paid in the LTC state i;

b_{a} rate of time-continuous benefits paid in the autonomy state a;

c_{ad} benefit paid in case of death if the policyholder occupies the autonomy state a;

c_{id} benefit paid in case of death if the policyholder occupies the LTC state i;

c_{ai} benefit paid when the policyholder enters the LTC state i.

For premiums, we denote as:

π_{i} rate of time-continuous premiums paid in the LTC state i;

π_{a} rate of time-continuous premiums paid in the autonomy state a.

All these quantities may be functions of time, i.e. $\pi_{\mathrm{a}} = \pi_{\mathrm{a}}(t)$ for instance. This allows the actuary to account for periods with no premiums due, for instance. Benefits and premiums in the LTC state may be functions of time and duration of stay in state i, i.e. $b_{\mathrm{i}} = b_{\mathrm{i}}(t,z)$ for instance, because of specific policy conditions.

In general $\pi_{\mathrm{i}} = 0$ but we keep here the possibility of charging premiums in the LTC state, for the sake of completeness. Clearly, in case premiums are charged while benefits are paid, the actuary can always reduce the benefits accordingly so that we assume that the identity

$$b_{\mathrm{i}}(t,z)\pi_{\mathrm{i}}(t,z) \quad = \quad b_{\mathrm{a}}(t)\pi_{\mathrm{a}}(t) = 0$$

holds for all t.

The equivalence principle then states that the expected present value of the premiums paid by the policyholder

$$\Pi \quad = \quad \int_0^{\omega_x} {}_tp_x^{\mathrm{aa}}\pi_{\mathrm{a}}(t)v(0,t)\mathrm{d}t + \int_0^{\omega_x} {}_tp_x^{\mathrm{aa}}\mu_{x+t}^{\mathrm{ai}} \left(\int_0^{\omega_x - t} {}_zp_{x+t;0}^{\mathrm{ii}}\pi_{\mathrm{i}}(t+z,z)v(0,t+z)\mathrm{d}z \right) \mathrm{d}t$$

matches the expected present value of the benefits comprised in the contract

$$\begin{aligned} B \quad = \quad & \int_0^{\omega_x} {}_tp_x^{\mathrm{aa}} b_{\mathrm{a}}(t)v(0,t)\mathrm{d}t \\ & + \int_0^{\omega_x} {}_tp_x^{\mathrm{aa}}\mu_{x+t}^{\mathrm{ai}} \left(\int_0^{\omega_x - t} {}_zp_{x+t;0}^{\mathrm{ii}} b_{\mathrm{i}}(t+z,z)v(0,t+z)\mathrm{d}z \right) \mathrm{d}t \\ & + \int_0^{\omega_x} v(0,t)\,{}_tp_x^{\mathrm{aa}}\mu_{x+t}^{\mathrm{ai}} c_{\mathrm{ai}}(t)\mathrm{d}t + \int_0^{\omega_x} v(0,t)\,{}_tp_x^{\mathrm{aa}}\mu_{x+t}^{\mathrm{ad}} c_{\mathrm{ad}}(t)\mathrm{d}t \\ & + \int_0^{\omega_x} {}_tp_x^{\mathrm{aa}}\mu_{x+t}^{\mathrm{ai}} \left(\int_0^{\omega_x - t} {}_zp_{x+t;0}^{\mathrm{ii}}\mu_{x+t+z;z}^{\mathrm{id}} c_{\mathrm{id}}(t+z,z)v(0,t+z)\mathrm{d}z \right) \mathrm{d}t, \end{aligned}$$

that is, the equality

$$\Pi \quad = \quad B$$

has to hold at policy issue. To make the age x at policy issue visible, we sometimes write Π_x for the single premium $\Pi = B$. The premium rates $\pi_{\mathrm{a}}(\cdot)$ and $\pi_{\mathrm{i}}(\cdot)$ are then set in such a way that the equivalence principle is fulfilled.

6.6 Generalized annuities

Henceforth, several generalized annuity values will be useful, so that we give them specific notations. Precisely, we consider actuarial values (i.e. expected present values) of the following time-continuous annuities:

$$\begin{aligned}
\overline{a}^{\mathrm{aa}}_{x+t} &= \int_0^{\omega_x-t} {}_sp^{\mathrm{aa}}_{x+t} v(t,t+s)\mathrm{d}s \\
\overline{a}^{\mathrm{ai}}_{x+t} &= \int_0^{\omega_x-t} {}_sp^{\mathrm{ai}}_{x+t} v(t,t+s)\mathrm{d}s \\
\overline{a}^{\mathrm{ii}}_{x+t;z} &= \int_0^{\omega_x-t} {}_sp^{\mathrm{ii}}_{x+t;z} v(t,t+s)\mathrm{d}s.
\end{aligned}$$

In case of temporary annuities, with payments limited to n years, the symbol "$;n\rceil$" is added after age $x+t$, like in

$$\overline{a}^{\mathrm{aa}}_{x+t;n\rceil} = \int_0^{n} {}_sp^{\mathrm{aa}}_{x+t} v(t,t+s)\mathrm{d}s.$$

Often, policy conditions specify a constant rate of premium payable as long as the insured is in state a. The single premium Π is then easily converted into the constant rate of premium π_{a} payable continuously in state a:

$$\pi_{\mathrm{a}} = \frac{\Pi}{\overline{a}^{\mathrm{aa}}_{x}}$$

if premium payment is lifelong, or

$$\pi_{\mathrm{a}} = \frac{\Pi}{\overline{a}^{\mathrm{aa}}_{x;n\rceil}}$$

if premium payment is temporary, limited to n years.

When the transition intensities are piecewise constant, these annuity values can be calculated explicitly because the integrals admit analytical solutions. The idea is to proceed as follows. For integer age x,

$$\begin{aligned}
\overline{a}^{\mathrm{aa}}_{x} &= \int_0^{\omega_x} {}_tp^{\mathrm{aa}}_{x} v(0,t)\mathrm{d}t \\
&= \int_0^{\omega_x} \exp\left(-\int_0^t \mu^{\mathrm{a}\bullet}_{x+s}\mathrm{d}s\right)\exp(-\delta t)\mathrm{d}t \\
&= \int_0^1 \exp(-t\mu^{\mathrm{a}\bullet}_{x}-t\delta)\mathrm{d}t + \exp(-\mu^{\mathrm{a}\bullet}_{x})\int_1^2 \exp(-(t-1)\mu^{\mathrm{a}\bullet}_{x+1}-t\delta)\mathrm{d}t \\
&\quad + \exp(-\mu^{\mathrm{a}\bullet}_{x}-\mu^{\mathrm{a}\bullet}_{x+1})\int_2^3 \exp(-(t-2)\mu^{\mathrm{a}\bullet}_{x+2}-t\delta)\mathrm{d}t + \ldots
\end{aligned}$$

Each of these integrals now admits an analytical expression so that we obtain

$$\begin{aligned}\overline{a}_x^{\mathrm{aa}} &= \frac{1-\exp(-\delta-\mu_x^{\mathrm{a}\bullet})}{\delta+\mu_x^{\mathrm{a}\bullet}} \\ &\quad + \sum_{j=1}^{\omega_x-1} \exp\left(-\sum_{k=0}^{j-1}\mu_{x+k}^{\mathrm{a}\bullet} - j\delta\right) \frac{1-\exp(-\delta-\mu_{x+j}^{\mathrm{a}\bullet})}{\delta+\mu_{x+j}^{\mathrm{a}\bullet}}.\end{aligned}$$

Proceeding in a similar way, we get

$$\begin{aligned}\overline{a}_{x;0}^{\mathrm{ii}} &= \int_0^{\omega_x} {}_tp_{x;0}^{\mathrm{ii}} v(0,t)\mathrm{d}t \\ &= \frac{1-\exp(-\delta-\widetilde{\mu}(x,0))}{\delta+\widetilde{\mu}(x,0)} \\ &\quad + \sum_{j=1}^{\omega_x-1} \exp\left(-\sum_{k=0}^{j-1}\widetilde{\mu}(x,k) - j\delta\right) \frac{1-\exp(-\delta-\widetilde{\mu}(x,j))}{\delta+\widetilde{\mu}(x,j)}.\end{aligned}$$

6.7 Generalized life insurances

The actuarial values of a unit lump sum paid in case of a transition, depending on the initial state, are given by

$$\begin{aligned}\overline{A}_{x+t}^{\mathrm{a;a\to i}} &= \int_0^{\omega_x-t} v(t,t+s)\,{}_sp_{x+t}^{\mathrm{aa}}\mu_{x+t+s}^{\mathrm{ai}}\mathrm{d}s \\ \overline{A}_{x+t}^{\mathrm{a;a\to d}} &= \int_0^{\omega_x-t} v(t,t+s)\,{}_sp_{x+t}^{\mathrm{aa}}\mu_{x+t+s}^{\mathrm{ad}}\mathrm{d}s \\ \overline{A}_{x+t}^{\mathrm{a;i\to d}} &= \int_0^{\omega_x-t} {}_sp_{x+t}^{\mathrm{aa}}\mu_{x+t+s}^{\mathrm{ai}}\left(\int_0^{\omega_x-t-s} {}_zp_{x+t+s;0}^{\mathrm{ii}}\mu_{x+t+s+z;z}^{\mathrm{id}} v(t,t+s+z)\mathrm{d}z\right)\mathrm{d}s \\ \overline{A}_{x+t;z}^{\mathrm{i;i\to d}} &= \int_0^{\omega_x-t} v(t,t+s)\,{}_sp_{x+t;z}^{\mathrm{ii}}\mu_{x+t+s;z+s}^{\mathrm{id}}\mathrm{d}s.\end{aligned}$$

In case of temporary benefits limited to n years, the symbol "$;n\rceil$" is added after age $x+t$, like

$$\overline{A}_{x+t;n\rceil}^{\mathrm{a;a\to i}} = \int_0^n v(t,t+s)\,{}_sp_{x+t}^{\mathrm{aa}}\mu_{x+t+s}^{\mathrm{ai}}\mathrm{d}s.$$

The transition has to occur within the next n years to get the insurance benefit.

When transition intensities are piecewise constant, we get

$$\begin{aligned}\overline{A}_x^{\mathrm{a;a\to d}} &= \int_0^{\omega_x} v(0,t)\,{}_tp_x^{\mathrm{aa}}\mu_{x+t}^{\mathrm{ad}}\mathrm{d}t \\ &= \mu_x^{\mathrm{ad}}\frac{1-\exp(-\delta-\mu_x^{\mathrm{a}\bullet})}{\delta+\mu_x^{\mathrm{a}\bullet}} \\ &\quad + \sum_{j=1}^{\omega_x-1}\mu_{x+j}^{\mathrm{ad}}\exp\left(-\sum_{k=0}^{j-1}\mu_{x+k}^{\mathrm{a}\bullet} - j\delta\right)\frac{1-\exp(-\delta-\mu_{x+j}^{\mathrm{a}\bullet})}{\mu_{x+j}^{\mathrm{a}\bullet}+\delta}\end{aligned}$$

with a similar expression for $\overline{A}_x^{\mathrm{a;a\to i}}$. For a unit death benefit granted to an individual who just entered the LTC state, we have

$$\begin{aligned}
\overline{A}_{x;0}^{\mathrm{i;i\to d}} &= \int_0^{\omega_x} {}_zp_{x;0}^{\mathrm{ii}}\mu_{x+z;z}^{\mathrm{id}}v(0,z)\mathrm{d}z \\
&= \widetilde{\mu}(x,0)\frac{1-\exp(-\delta-\widetilde{\mu}(x,0))}{\delta+\widetilde{\mu}(x,0)} \\
&\quad + \sum_{j=1}^{\omega_x-1}\widetilde{\mu}(x,j)\exp\left(-\sum_{k=0}^{j-1}\widetilde{\mu}(x,k)-j\delta\right)\frac{1-\exp(-\delta-\widetilde{\mu}(x,j))}{\delta+\widetilde{\mu}(x,j)}.
\end{aligned}$$

In case of a death benefit in the LTC state, granted to an autonomous individual, we have

$$\begin{aligned}
\overline{A}_x^{\mathrm{a;i\to d}} &= \int_0^{\omega_x} {}_tp_x^{\mathrm{aa}}\mu_{x+t}^{\mathrm{ai}}\left(\int_0^{\omega_x-t} {}_zp_{x+t;0}^{\mathrm{ii}}\mu_{x+t+z;z}^{\mathrm{id}}v(0,t+z)\mathrm{d}z\right)\mathrm{d}t \\
&= \mu_x^{\mathrm{ai}}\overline{A}_{x;0}^{\mathrm{i;i\to d}}\frac{1-\exp(-\delta-\mu_x^{\mathrm{a\bullet}})}{\delta+\mu_x^{\mathrm{a\bullet}}} \\
&\quad + \sum_{j=1}^{\omega_x-1}\mu_{x+j}^{\mathrm{ai}}\overline{A}_{x+j;0}^{\mathrm{i;i\to d}}\exp\left(-\sum_{k=0}^{j-1}\mu_{x+k}^{\mathrm{a\bullet}}-j\delta\right)\frac{1-\exp(-\delta-\mu_{x+j}^{\mathrm{a\bullet}})}{\delta+\mu_{x+j}^{\mathrm{a\bullet}}}.
\end{aligned}$$

6.8 Some specific conditions

Several policy conditions can be included in LTC insurance products. In this section we only address duration-related conditions, i.e. policy conditions which either define the coverage period or the benefit payment period following the claim, that is, the inception of the LTC need.

6.8.1 Insured period

The insured period (or "coverage" period) is the time interval during which the insurance cover operates, in the sense that a benefit is payable only if the claim time belongs to this interval. In principle, the insured period begins at policy issue, and ends at policy termination. In LTC policies, given the purpose of the benefits, it is reasonable to assume a lifelong insured period. However some restrictions to the insured period may follow from specific policy conditions.

6.8.2 Waiting, or elimination period

The waiting period (or "elimination" period) is the period following the policy issue during which the insurance cover is not yet operating for sickness-related claims (loss of autonomy due to an accident is generally covered without limitation, from the beginning of the insured period). Different waiting periods can be applied according to the category of sickness. The waiting period aims at limiting the effects of adverse selection, in

particular because of pre-existing insured's health conditions. It is worth noticing that, although the term waiting period is widely adopted, this time interval is sometimes called the "probationary" period (for instance in the US), while the term waiting period is used synonymously with "deferred" period (see below).

In case there is a transition to state i before the end of the waiting period (of duration w, say), the insurer may pay back the premium charged so far, i.e.

$$c_{\mathrm{ai}}(t) = \int_0^t \pi_{\mathrm{a}}(s)\mathrm{d}s \text{ for } t < w$$

when nominal amounts are reimbursed. We note that $c_{\mathrm{ai}}(t)$ constitutes a counter-insurance benefit.

6.8.3 Deferred period

In many policies the benefit is not payable until the LTC need has lasted a certain minimum period called the deferred period. This policy condition has a two-fold purpose:

- to reduce the cost and hence the premium of the LTC insurance product; premium reduction can be particularly significant because of the high mortality immediately following the loss of autonomy;
- to ascertain the permanent character of the disease which implies the LTC need (provided that LTC benefits are only paid in the case of permanent disability, as assumed in our model).

6.9 Premium formulas for some LTC insurance products

We will refer here to the following products:

1. the stand-alone LTC cover;
2. the enhanced pension, or life care annuity;
3. a package of LTC and lifetime-related benefits;
4. the whole-life insurance with LTC acceleration benefit;
5. an LTC package combining a whole life insurance product comprising a surrender option in case of loss of autonomy, offsetting the financial impact of the deferred period of the LTC annuity.

Formulae for the single premiums of the above products are provided hereafter.

We note that products 2 to 5 constitute special types of insurance packages, or "combined products". From the insurer's perspective, a combined product may be profitable even if one of its components is not profitable. In addition, a combined product may be less risky if it includes some internal hedging mechanism. We refer the reader e.g. to Pitacco (2016) for several examples.

6.9.1 Stand-alone LTC cover

The benefit consists in a time-continuous annuity, continuously paid at constant rate b_{i}, while the insured is in state i. In the case of no time restriction (that is, in the base case), the single premium is given by:

$$\Pi = b_{\mathrm{i}} \overline{a}_x^{\mathrm{ai}}.$$

Notice that Π can be alternatively rewritten as

$$\Pi = b_{\mathrm{i}} \int_0^{\omega_x} {}_t p_x^{\mathrm{aa}} \mu_{x+t}^{\mathrm{ai}} v(0,t) \overline{a}_{x+t;0}^{\mathrm{ii}} \mathrm{d}t$$

which makes explicit the time t of entry in the LTC state. This second formula is useful when policy conditions state some duration-related restrictions, as shown next.

In case policy conditions specify a waiting (or elimination) period w, the single premium becomes

$$\Pi = b_{\mathrm{i}} \int_w^{\omega_x} {}_t p_x^{\mathrm{aa}} \mu_{x+t}^{\mathrm{ai}} v(0,t) \overline{a}_{x+t;0}^{\mathrm{ii}} \mathrm{d}t.$$

In case policy conditions specify a deferred period d, we get

$$\Pi = b_{\mathrm{i}} \int_0^{\omega_x} {}_t p_x^{\mathrm{aa}} \mu_{x+t\,d}^{\mathrm{ai}} p_{x+t;0}^{\mathrm{ii}} v(0,t+d) \overline{a}_{x+t+d;d}^{\mathrm{ii}} \mathrm{d}t.$$

Finally, in case policy conditions specify both a waiting period w and a deferred period d, the single premium is given by

$$\Pi = b_{\mathrm{i}} \int_w^{\omega_x} {}_t p_x^{\mathrm{aa}} \mu_{x+t\,d}^{\mathrm{ai}} p_{x+t;0}^{\mathrm{ii}} v(0,t+d) \overline{a}_{x+t+d;d}^{\mathrm{ii}} \mathrm{d}t.$$

If transition intensities are piecewise constant, then the single premium of a stand-alone LTC cover can be computed as follows:

$$\begin{aligned}
\Pi &= b_{\mathrm{i}} \overline{a}_x^{\mathrm{ai}} \\
&= b_{\mathrm{i}} \int_0^{\omega_x} \exp\left(-\delta t - \int_0^t \mu_{x+s}^{\mathrm{a}\bullet} \mathrm{d}s \right) \mu_{x+t}^{\mathrm{ai}} \overline{a}_{x+t;0}^{\mathrm{ii}} \mathrm{d}t \\
&= b_{\mathrm{i}} \mu_x^{\mathrm{ai}} \overline{a}_{x;0}^{\mathrm{ii}} \frac{1 - \exp(-\delta - \mu_x^{\mathrm{a}\bullet})}{\delta + \mu_x^{\mathrm{a}\bullet}} \\
&\quad + b_{\mathrm{i}} \sum_{j=1}^{\omega_x - 1} \mu_{x+j}^{\mathrm{ai}} \overline{a}_{x+j;0}^{\mathrm{ii}} \exp\left(-\sum_{k=0}^{j-1} \mu_{x+k}^{\mathrm{a}\bullet} - j\delta \right) \frac{1 - \exp(-\delta - \mu_{x+j}^{\mathrm{a}\bullet})}{\delta + \mu_{x+j}^{\mathrm{a}\bullet}}.
\end{aligned}$$

The values of Π are displayed as a function of age x at policy issue in Figure 6.5 with $b_{\mathrm{i}} = 12,000$ (i.e. a monthly payment of 1,000). We can see there that the amount of the single premium increases rapidly until age 80, where it stabilizes due to effect of high mortality (as this product does not comprise any benefit in case of death).

Let us now introduce a waiting period and a deferred period. In case of a waiting period of one year (i.e. $w = 1$), the premium becomes

$$\begin{aligned}\Pi &= b_{\mathrm{i}} \int_1^{\omega_x} \exp\left(-\delta t - \int_0^t \mu^{\mathrm{a}\bullet}_{x+s}\mathrm{d}s\right) \mu^{\mathrm{ai}}_{x+t}\, \overline{a}^{\mathrm{ii}}_{x+t:0}\mathrm{d}t \\ &= b_{\mathrm{i}} \sum_{j=1}^{\omega_x-1} \mu^{\mathrm{ai}}_{x+j}\, \overline{a}^{\mathrm{ii}}_{x+j:0} \exp\left(-\sum_{k=0}^{j-1} \mu^{\mathrm{a}\bullet}_{x+k} - j\delta\right) \frac{1-\exp(-\delta-\mu^{\mathrm{a}\bullet}_{x+j})}{\delta+\mu^{\mathrm{a}\bullet}_{x+j}}.\end{aligned}$$

Including a deferred period $d \leq 1$, we get

$$\begin{aligned}\Pi_x &= b_{\mathrm{i}} \int_0^{\omega_x} \exp\left(-\delta(t+d) - \int_0^t \mu^{\mathrm{a}\bullet}_{x+s}\mathrm{d}s\right) \mu^{\mathrm{ai}}_{x+t} \exp\left(-\int_0^d \mu^{\mathrm{id}}_{x+t+z;z}\mathrm{d}z\right) \overline{a}^{\mathrm{ii}}_{x+t+d;d}\mathrm{d}t \\ &= b_{\mathrm{i}} \overline{a}^{\mathrm{ii}}_{x+d;d} \exp\Big(-d(\delta+\widetilde{\mu}(x,0))\Big) \mu^{\mathrm{ai}}_x \frac{1-\exp(-\delta-\mu^{\mathrm{a}\bullet}_x)}{\delta+\mu^{\mathrm{a}\bullet}_x} \\ &\quad + b_{\mathrm{i}} \sum_{j=1}^{\omega_x-1} \overline{a}^{\mathrm{ii}}_{x+d+j;d} \mu^{\mathrm{ai}}_{x+j} \exp\left(-d(\delta d+\widetilde{\mu}(x+j,0)) - \sum_{k=0}^{j-1} \mu^{\mathrm{a}\bullet}_{x+k} - j\delta\right) \frac{1-\exp(-\delta-\mu^{\mathrm{a}\bullet}_{x+j})}{\delta+\mu^{\mathrm{a}\bullet}_{x+j}}.\end{aligned}$$

The diminishing effect of the inclusion of a waiting period and of a deferred period in policy conditions is illustrated in Figure 6.6. Whereas the waiting period moderately decreases the amount of the single premium (the reduction getting nevertheless larger as the age at policy issue increases), the deferred period greatly reduces the single premium because of the high mortality just after the loss of autonomy. The impact of varying the deferred period is illustrated on Figure 6.7. We can see there that the higher the deferred period, the lower the single premium, as expected.

6.9.2 Enhanced pension, or life care annuity

Enhanced pensions, or life care annuities, are life annuity products in which the LTC benefit is defined in terms of an uplift with respect to the basic pension. Benefits are then defined as follows. The life annuity, that is, the basic pension, is payable continuously at rate b_{a} in state a. The LTC annuity is payable at rate b_{i} in state i, with $b_{\mathrm{i}} > b_{\mathrm{a}}$. The integration of life annuity and LTC cover into a single product is expected to broaden the population that can be insured, as those individuals with high risk on one component of the package are generally better risks on the other one (see, e.g., Brown and Warshawsky, 2013).

The single premium is then given by

$$\begin{aligned}\Pi &= b_{\mathrm{a}}\overline{a}^{\mathrm{aa}}_x + b_{\mathrm{i}}\overline{a}^{\mathrm{ai}}_x \\ &= b_{\mathrm{a}}\overline{a}^{\mathrm{a}}_x + (b_{\mathrm{i}}-b_{\mathrm{a}})\overline{a}^{\mathrm{ai}}_x\end{aligned}$$

where $b_{\mathrm{i}} - b_{\mathrm{a}}$ is the uplift amount, and

$$\overline{a}^{\mathrm{a}}_x = \overline{a}^{\mathrm{aa}}_x + \overline{a}^{\mathrm{ai}}_x$$

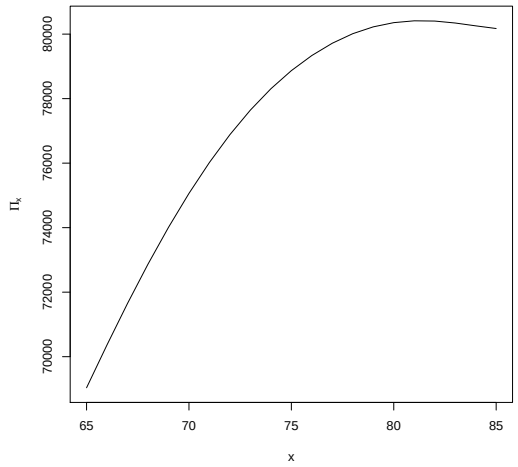

Figure 6.5: Single premium Π_x as a function of age x at policy issue of a stand-alone LTC cover with $b_{\mathrm{i}} = 12,000$ without limitations.

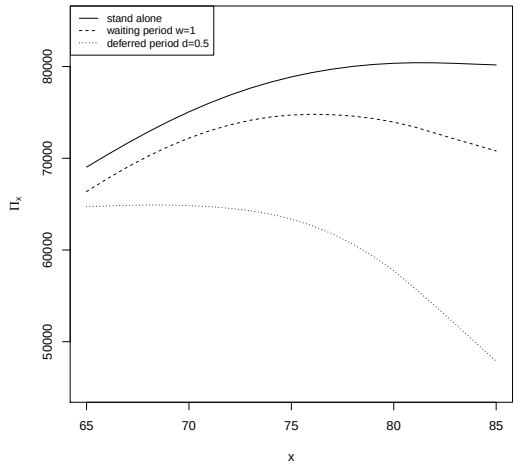

Figure 6.6: Impact of the waiting period $w = 1$ and of the deferred period $d = 0.5$ on the single premium Π_x of a stand-alone LTC cover, as a function of age x at policy issue with $b_{\mathrm{i}} = 12,000$.

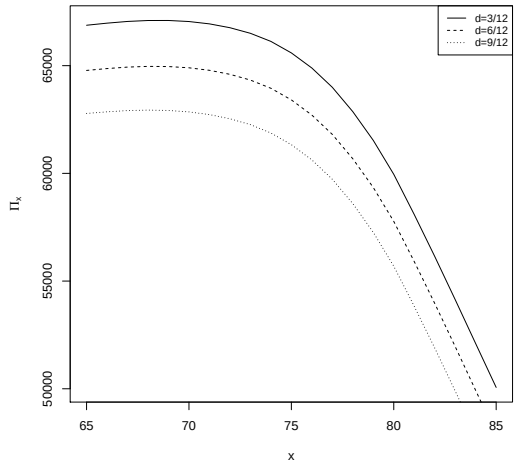

Figure 6.7: Impact of increasing the deferred period d from 3 to 9 months on the single premium Π_x of a stand-alone LTC cover, as a function of age x at policy issue with $b_{\mathrm{i}} = 12,000$.

is the price of a regular life annuity sold to an autonomous individual. If $b_{\mathrm{i}} = b_{\mathrm{a}}$, the product comes down to a usual life annuity sold to an autonomous individual. Henceforth, we set $b_{\mathrm{i}} = 2b_{\mathrm{a}}$.

The single premiums of the enhanced pension are displayed in Figure 6.8 as a function of age at policy issue for $b_{\mathrm{a}} = 12,000$. Clearly, Π_x now decreases with age at policy issue. Notice that in this graph, ages at entry up to 85 are considered. If x is greater than 75, an old-age pension (i.e. not a standard one) is involved in the considered package.

6.9.3 Package of LTC and lifetime-related benefits

An insurance package can include LTC benefits combined with various lifetime-related benefits, i.e. benefits only depending on insured's survival and death. We consider the package in which the following benefits are included:

- a deferred life annuity payable at constant rate b_{a}, paid while the insured is in state a (the deferment period is denoted as n);
- a LTC annuity payable at a rate b_{i} paid while the insured is in state i;
- death benefits of amount

$$c_{\mathrm{ad}} = c_{\mathrm{id}} = c_{\mathrm{d}}.$$

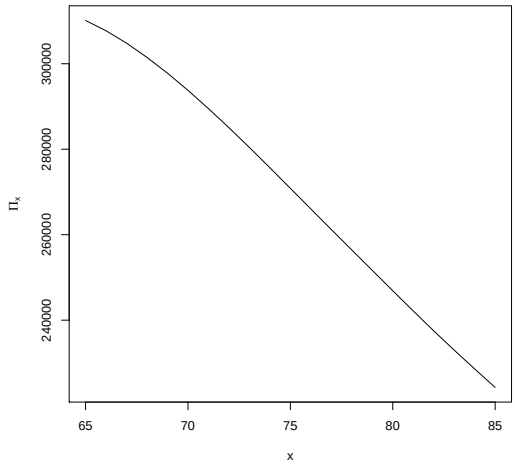

Figure 6.8: Single premium Π_x of an enhanced pension as a function of age x at policy issue, with $b_i = 2b_a = 24,000$.

The single premium is then given by

$$\Pi = b_{\mathrm{a}} v(0,n)\, {}_n p_x^{\mathrm{aa}} \overline{a}_{x+n}^{\mathrm{aa}} + b_{\mathrm{i}} \overline{a}_x^{\mathrm{ai}} + c_{\mathrm{d}} \overline{A}_x^{\mathrm{a}}$$

where

$$\overline{A}_x^{\mathrm{a}} = \overline{A}_x^{\mathrm{a;a}\to\mathrm{d}} + \overline{A}_x^{\mathrm{a;i}\to\mathrm{d}}$$

is the price of a whole life insurance sold to an individual in autonomy and

$$\begin{aligned} {}_n p_x^{\mathrm{aa}} &= \exp\left(-\int_0^n \mu_{x+t}^{\mathrm{a}\bullet} \mathrm{d}t \right) \\ &= \exp\left(-\sum_{j=0}^{n-1} \mu_{x+j}^{\mathrm{a}\bullet} \right). \end{aligned}$$

According to an alternative definition, the death benefits c_{ad} and c_{id} are given by the difference (if positive) between a stated amount c and the amount totally paid as deferred life annuity and/or LTC annuity.

6.9.4 Whole-life insurance with LTC acceleration benefit

LTC benefits can be added as a rider to a whole-life insurance policy. In particular, the LTC annuity benefit can (totally or partially) be financed by "accelerating" the payment of (part of) the death benefit. Specifically, let c_{ad} be the amount of death benefit for a

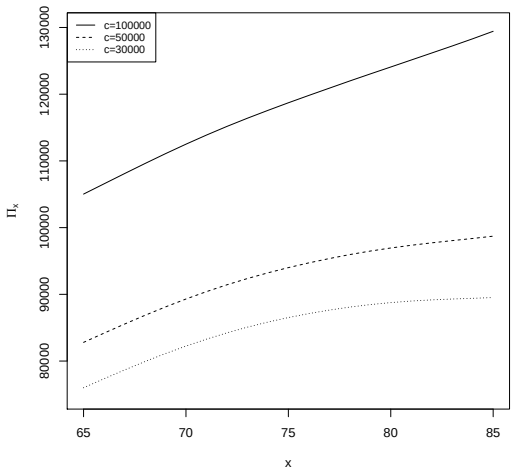

Figure 6.9: Single premium of the whole-life insurance cover with LTC acceleration benefit, as a function of age x at policy issue, for $b_{\mathrm{i}} = 12,000$ and $c_{\mathrm{i}} = c_{\mathrm{a}} = c$.

policyholder in state a. The LTC annuity is payable continuously at rate b_{i}. The amount of death benefit for an insured in state i dying after having spent a duration z in state i is given by

$$c_{\mathrm{id}}(t,z) = \max\{c_{\mathrm{ad}} - b_{\mathrm{i}} z, 0\} = (c_{\mathrm{ad}} - b_{\mathrm{i}} z)_+.$$

In case only the death benefit (or part of the death benefit) is converted into an LTC annuity, the single premium is given by:

$$\Pi = c_{\mathrm{ad}} \overline{A}_x^{\mathrm{a;a\to d}} + b_{\mathrm{i}} \overline{a}^{\mathrm{ai}}_{x;\lceil c_{\mathrm{ad}}/b_{\mathrm{i}} \rceil} + \int_0^{\omega_x} {}_t p_x^{\mathrm{aa}} \mu_{x+t}^{\mathrm{ai}} \left(\int_0^{c_{\mathrm{ad}}/b_{\mathrm{i}}} (c_{\mathrm{ad}} - b_{\mathrm{i}} z)_z p^{\mathrm{ii}}_{x+t;0} \mu^{\mathrm{id}}_{x+t+z;z} v(0,t+z) \mathrm{d}z \right) \mathrm{d}t.$$

In case the LTC annuity is paid until death, with the death benefit decreased accordingly until possible exhaustion, the single premium is given by:

$$\Pi = c_{\mathrm{ad}} \overline{A}_x^{\mathrm{a;a\to d}} + b_{\mathrm{i}} \overline{a}^{\mathrm{ai}}_x + \int_0^{\omega_x} {}_t p_x^{\mathrm{aa}} \mu_{x+t}^{\mathrm{ai}} \left(\int_0^{c_{\mathrm{ad}}/b_{\mathrm{i}}} (c_{\mathrm{ad}} - b_{\mathrm{i}} z)_z p^{\mathrm{ii}}_{x+t;0} \mu^{\mathrm{id}}_{x+t+z;z} v(0,t+z) \mathrm{d}z \right) \mathrm{d}t.$$

Of course, the second arrangement yields a single premium greater than the first one.

Figure 6.9 displays the single premium of this second arrangement. We can see there that the premium increases with age x at policy issue, this increase being steeper for higher amount of death benefit.

6.9.5 LTC package with a whole-life insurance offsetting the deferred period

This package consists of a whole-life insurance coverage together with a LTC annuity with monthly payment m, subject to a deferred period. The length of the deferred period depends on policyholder's age x at policy issue; henceforth, we denote it as d_x. The single premium of the package is $m \times d_x$, where the policyholder selects the desired value of m and the insurer's tariff gives d_x corresponding to policyholder's age x. This is also the amount of benefit comprised in the whole-life insurance cover, which authorizes surrender at entry in the LTC state and thus offsets the financial impact of the deferred period.

Prospects are told that they get their premium $m \times d_x$ back in any case, either at death, at loss of autonomy, or at surrender in state a (but the LTC cover is then automatically cancelled). In case of entry in the LTC state, policyholders can use this amount as benefits during the deferred period (as $m \times d_x$ is precisely the amount needed for the d_x months during which no LTC benefits are paid). Thus, it seems that the LTC cover comes for free, which is particularly attractive from the policyholder's point of view. Let us stress that all benefits are specified in absolute terms (and are not re-evaluated to compensate for inflation) and that the product is sold at relatively young ages (before retirement, in any case).

The product is sold as a combination of a life insurance contract and a LTC cover. Depending on the country, these two products may fall under different lines of business and must then be managed separately. Both products are assumed to have the same technical interest rate δ.

For the whole life insurance cover, the single premium Π_{wl} is also the amount of benefit paid in case of death in autonomy, in case of loss of autonomy, or in case of surrender in autonomy. The policyholder is allowed to surrender at any time (but this automatically cancels the loss of autonomy cover if the policyholder is in state a). The optimal behavior thus consists in surrendering the whole-life insurance contract at entry in state i so that the surrender value can be used as LTC benefits during the deferred period of length d_x.

According to the policy conditions, the insurer charges expenses proportional to the reserve at rate δ on the whole life insurance product. These expenses serve as premiums for the LTC cover (the whole package thus requires a sufficiently high interest rate to be effective). This results in a zero interest rate for the whole-life insurance contract, as deduced from Thiele equation: the calculations can be carried out for the life insurance component as if the technical interest rate was set to zero. For the whole-life insurance component, if the state s denotes surrender (so that we implicitly work here with a 4th state), the equivalence equation

$$\begin{aligned} \Pi_{\text{wl}} &= \int_0^{\omega_x} {}_t p_x^{\text{aa}} (\mu_{x+t}^{\text{ai}} + \mu_{x+t}^{\text{as}} + \mu_{x+t}^{\text{ad}}) \Pi_{\text{wl}} \mathrm{d}t \\ &= \Pi_{\text{wl}} \end{aligned}$$

is obviously valid whatever Π_{wl}. Here, we take $\Pi_{\text{wl}} = m d_x$ which is the full price of the package. The LTC component is then paid by the expenses charged on the whole-life

insurance cover, that is,

$$\pi_{\mathrm{a}}(t) = \delta m d_x.$$

This is because the reserve V_{wl} of the whole-life insurance contract is just the expected present value of future benefits (because the contract stipulates a single premium, see the next section for the formal definition of the reserve), i.e.

$$\begin{aligned} V_{\mathrm{wl}}(t) &= \int_0^{\omega_x - t} {}_s p_{x+t}^{\mathrm{aa}} (\mu_{x+t+s}^{\mathrm{ai}} + \mu_{x+t+s}^{\mathrm{as}} + \mu_{x+t+s}^{\mathrm{ad}}) \Pi_{\mathrm{wl}} \mathrm{d}s \\ &= \Pi_{\mathrm{wl}} \\ &= m d_x. \end{aligned}$$

For the loss of autonomy component, as premiums are paid continuously in state a at rate $\delta m d_x$, the length of the deferred period is the unique solution of

$$\begin{aligned} &\textstyle\int_0^{\omega_x} {}_t p_x^{\mathrm{aa}} \exp(-\delta t) \delta m d \mathrm{d}t \\ &\quad = 12m \int_0^{\omega_x} {}_t p_x^{\mathrm{aa}} \mu_{x+t\,d/12}^{\mathrm{ai}} p_{x+t;0}^{\mathrm{ii}} \exp\big(-\delta(t + d/12)\big) \bar{a}_{x+t+d/12;d/12}^{\mathrm{ii}} \mathrm{d}t. \end{aligned} \tag{6.9.1}$$

The uniqueness of the solution of (6.9.1) results from the following argument. The left-hand side of the equivalence relation (6.9.1) increases linearly in d, starting from the origin. The right-hand side decreases in d, starting from the strictly positive value $12m\bar{a}_x^{\mathrm{ai}}$. Therefore, by continuity, there must be a unique value fulfilling the equivalence constraint (6.9.1). Notice that m cancels on both sides of (6.9.1) so that there is a unique value of d for each age x at policy issue.

Assume that the policyholder aged 65 has selected the desired value of $m = 1{,}000$. We then have to find the unique value of d solving the equivalence relation (6.9.1). The left-hand side of the equivalence relation (6.9.1) increases linearly in d, starting from the origin, as shown in the top panel of Figure 6.10. The right-hand side decreases in d, starting from a strictly positive value, as shown in the bottom panel of Figure 6.10. The unique value of d fulfilling the equivalence constraint (6.9.1) is shown graphically on Figure 6.11. This results in $d_{65} = 58$ months for a yearly interest rate of 3% and $d_{65} = 33$ months for a yearly interest rate of 5%. In order to be conservative, d_{65} is chosen as the first integer for which the function is negative. As expected, d_{65} decreases with the interest rate, because a higher interest rate means a higher premium and a higher discounting for the LTC component.

6.10 Reserves

6.10.1 Principle

LTC contracts are generally lifelong with a level premium fixed at contract initiation so that the annual paid amount does not vary during the contract. This constant premium or level premium depends on the underwriting age x.

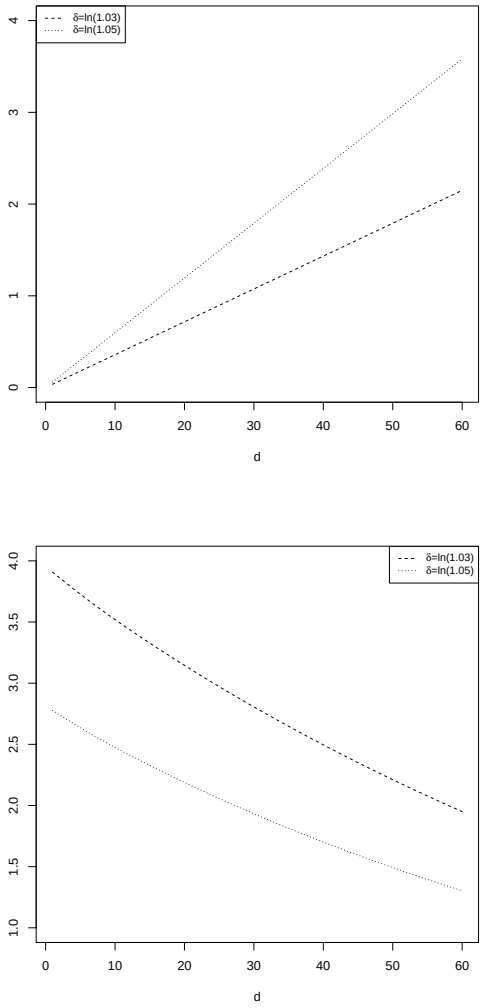

Figure 6.10: Left-hand side of the equivalence relation (6.9.1) in the top panel, right-hand side of the equivalence relation (6.9.1) in the bottom panel, age $x = 65$ at policy issue for a yearly interest rate of 3% or 5%.

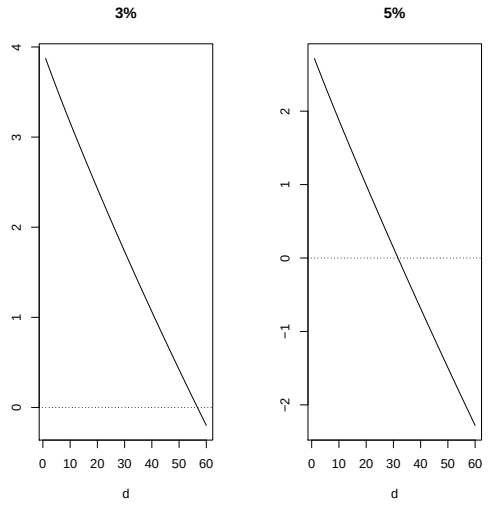

Figure 6.11: Graphical search for d_{65} for a yearly interest rate of 3% or 5%.

As it can be seen on Figure 6.12, the annual risk premium (i.e. annual expected claim amount for an active individual) is an increasing function of age except at very advanced ages. Therefore surpluses are constituted in the first part of the contract as level premiums exceed annual risk premiums. This surplus is called reserve and is kept aside to meet future needs.

In the case a single premium is paid at policy issue, a reserve must immediately be kept aside, and then "used" throughout the whole policy duration to meet the insurer's expected costs.

By status the insurer should have available a reserve at any time. It is defined prospectively as the actuarial value of future benefits less the actuarial value of future premiums (and, in the case of a single premium, is simply defined as the actuarial value of future benefits). Therefore, it depends on the state occupied by the policyholder at the date of calculation. In LTC insurance, we distinguish a reserve in state a at time t, henceforth denoted as V_t^{a}, and a reserve in state i at time t, with autonomy lost at time $t-z$, henceforth denoted as $V_{t;z}^{\mathrm{i}}$. Of course, there is no need to define a reserve in state d as policyholder's death automatically terminates the contract.

The equivalence principle states that, at policy issue the expected present value Π of the premiums paid by the policyholder matches the expected present value B of the benefits comprised in the contract, i.e. $V_0^{\mathrm{a}} = 0$. This equivalence does no more hold in the course of the contract. The reserve is the amount needed to restore financial equilibrium at any

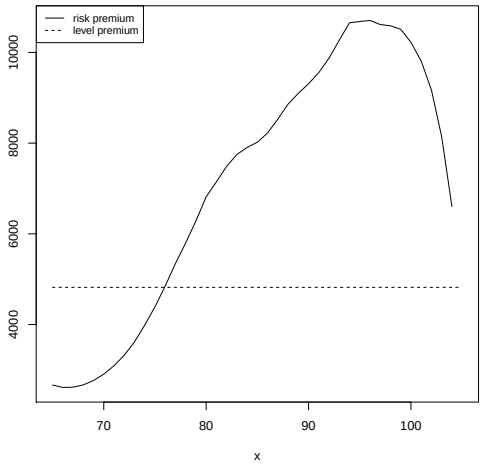

Figure 6.12: Annual risk premium by age, stand-alone LTC cover, and level premium π_{a} for a policyholder aged 65 at policy issue.

time $t > 0$. Specifically, let

$$\begin{aligned}\Pi_{\mathrm{a}}(t) &= \int_0^{\omega_x-t} {}_sp_{x+t}^{\mathrm{aa}}\pi_{\mathrm{a}}(t+s)v(t,t+s)\mathrm{d}s \\ &\quad + \int_0^{\omega_x-t} {}_sp_{x+t}^{\mathrm{aa}}\mu_{x+t+s}^{\mathrm{ai}}\left(\int_0^{\omega_x-t-s} {}_zp_{x+t+s;0}^{\mathrm{ii}}\pi_{\mathrm{i}}(t+s+z,z)v(t,t+s+z)\mathrm{d}z\right)\mathrm{d}s\end{aligned}$$

be the expected present value of the future premiums paid by an individual in state a at time t. Similarly, let

$$\begin{aligned}B_{\mathrm{a}}(t) &= \int_0^{\omega_x-t} {}_sp_{x+t}^{\mathrm{aa}}b_{\mathrm{a}}(t+s)v(t,t+s)\mathrm{d}s \\ &\quad + \int_0^{\omega_x-t} {}_sp_{x+t}^{\mathrm{aa}}\mu_{x+t+s}^{\mathrm{ai}}\left(\int_0^{\omega_x-t-s} {}_zp_{x+t+s;0}^{\mathrm{ii}}b_{\mathrm{i}}(t+s+z,z)v(t,t+s+z)\mathrm{d}z\right)\mathrm{d}s \\ &\quad + \int_0^{\omega_x-t} v(t,t+s){}_sp_{x+t}^{\mathrm{aa}}\mu_{x+t+s}^{\mathrm{ai}}c_{\mathrm{ai}}(t+s)\mathrm{d}s \\ &\quad + \int_0^{\omega_x-t} v(t,t+s){}_sp_{x+t}^{\mathrm{aa}}\mu_{x+t+s}^{\mathrm{ad}}c_{\mathrm{ad}}(t+s)\mathrm{d}s \\ &\quad + \int_0^{\omega_x-t} {}_sp_{x+t}^{\mathrm{aa}}\mu_{x+t+s}^{\mathrm{ai}}\left(\int_0^{\omega_x-t-s} {}_zp_{x+t+s;0}^{\mathrm{ii}}\mu_{x+t+s+z;z}^{\mathrm{id}}c_{\mathrm{id}}(t+s+z;z)v(t,t+s+z)\mathrm{d}z\right)\mathrm{d}s\end{aligned}$$

be the expected present value of the future benefits for an individual in state a at time t. Clearly, the quantities Π and B entering the equivalence principle are

$$\Pi = \Pi_{\mathrm{a}}(0) \text{ and } B = B_{\mathrm{a}}(0).$$

In case the waiting period w for LTC claims is not exhausted yet at reserve calculation, the second integral in $B_{\mathrm{a}}(t)$ does not start from 0 but from $w-t$ (the remaining part of the initial waiting period w specified in policy conditions). The deferred period is accounted for by appropriately defining the rate of benefit $b_{\mathrm{i}}(t+s+z,z)$, setting it to 0 as long as $z \leq d$.

The reserve in state a is then defined as

$$V_t^{\mathrm{a}} = B_{\mathrm{a}}(t) - \Pi_{\mathrm{a}}(t)$$

and represents the share of future benefits not covered by future premiums (so that the insurer must have accumulated this amount from past premiums). Adding the reserve to the expected present value of future benefits, the product is in financial equilibrium as

$$\Pi_{\mathrm{a}}(t) + V_t^{\mathrm{a}} = B_{\mathrm{a}}(t)$$

holds for all $t \geq 0$.

Let us now consider a policyholder in state i at time t, who entered the LTC state at time $t-z$. We assume that $t-z > w$ so that the LTC claim is not excluded because of the waiting period. If $t-z \leq w$ then the contract usually terminates and the insurer sometimes pays the total premiums paid so far, i.e. c_{ai} accumulates all the premiums paid in state a, until the transition to state i during the waiting period. For $t-z > w$, the reserve is given by

$$V_{t;z}^{\mathrm{i}} = B_{\mathrm{i}}(t;z) - \Pi_{\mathrm{i}}(t;z)$$

where

$$\Pi_{\mathrm{i}}(t;z) = \int_0^{\omega_x - t} {}_sp_{x+t;z}^{\mathrm{ii}} \pi_{\mathrm{i}}(t+s,z+s) v(t,t+s) \mathrm{d}s$$

and

$$\begin{aligned} B_{\mathrm{i}}(t;z) &= \int_0^{\omega_x - t} {}_sp_{x+t;z}^{\mathrm{ii}} b_{\mathrm{i}}(t+s,z+s) v(t,t+s) \mathrm{d}s \\ &\quad + \int_0^{\omega_x - t} v(t,t+s) {}_sp_{x+t;z}^{\mathrm{ii}} \mu_{x+t+s;z+s}^{\mathrm{id}} c_{\mathrm{id}}(t+s,z+s) \mathrm{d}s. \end{aligned}$$

In case policy conditions specify a deferred period d, the latter formula for $B_{\mathrm{i}}(t;z)$ is valid as long as $z > d$. For $z < d$, the policyholder must first spend an extra time $d-z$ in the LTC state before benefits start to be paid. The first integral appearing in $B_{\mathrm{i}}(t;z)$ then becomes

$${}_{d-z}p_{x+t;z}^{\mathrm{ii}} \int_0^{\omega_x - d} {}_sp_{x+t+d-z;d}^{\mathrm{ii}} b_{\mathrm{i}}(t+d-z+s,d+s) v(t,t+d-z+s) \mathrm{d}s.$$

6.10.2 Reserve formulas for some LTC insurance products

In this section, we assume that the premium is paid continuously, at constant rate π_{a}, as long as the policyholder stays in state a.

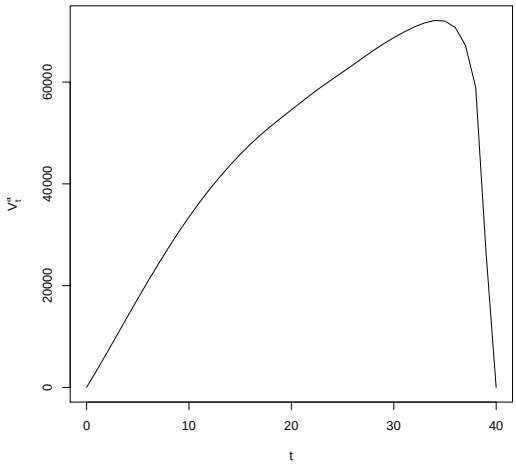

Figure 6.13: Evolution of the reserve V_t^{a} for the stand-alone LTC cover, as a function of time t for a policyholder aged 65 at policy issue with $b_{\mathrm{i}} = 12,000$.

Stand-alone LTC cover

The reserve at time t for an autonomous individual is equal to

$$V_t^{\mathrm{a}} = b_{\mathrm{i}} \overline{a}_{x+t}^{\mathrm{ai}} - \pi_{\mathrm{a}} \overline{a}_{x+t}^{\mathrm{aa}}$$

whereas the reserve for an individual in the LTC state at that time, who lost autonomy at time $t-z$, is equal to

$$V_{t;z}^{\mathrm{i}} = b_{\mathrm{i}} \overline{a}_{x+t;z}^{\mathrm{ii}}.$$

The reserve for an autonomous individual V_t^{a} is represented on Figure 6.13 as a function of t, for an initial age $x = 65$. The amount of reserve increases until the age of 100, before falling to 0 due to the high mortality risk.

Let us now examine the reserve $V_{t;z}^{\mathrm{i}}$ in the LTC state, as a function of z. Figure 6.14 displays the curve $z \mapsto V_{15;z}^{\mathrm{i}}$ for a policyholder aged 65 at policy issue (so that the age at reserve calculation is 80). We see that $z \mapsto V_{15;z}^{\mathrm{i}}$ first increases until the end of the first year spent in LTC (i.e. for $z \leq 1$) and then decreases. This results from the high mortality during the year following the entry in LTC state.

Enhanced pension

The reserve at time t for an autonomous individual is equal to

$$V_t^{\mathrm{a}} = b_{\mathrm{a}} \overline{a}_{x+t}^{\mathrm{aa}} + b_{\mathrm{i}} \overline{a}_{x+t}^{\mathrm{ai}} - \pi_{\mathrm{a}} \overline{a}_{x+t}^{\mathrm{aa}}$$

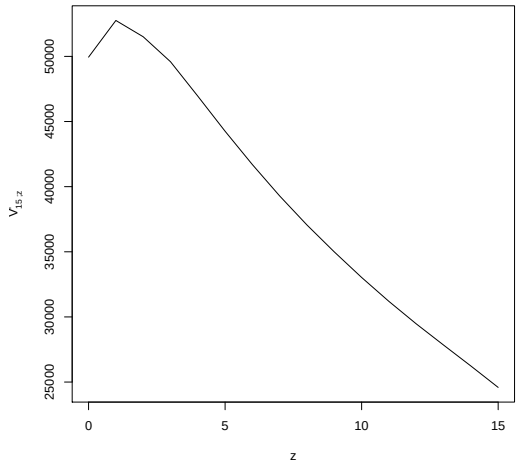

Figure 6.14: Evolution of the reserve $V^{\mathrm{i}}_{15;z}$ for the stand-alone LTC cover, as a function of the time z spent in the LTC state for a policyholder aged 65 at policy issue.

Considering an individual who is in the LTC state at time t, who entered that state at time $t-z$, the reserve is given by

$$V^{\mathrm{i}}_{t;z} = b_{\mathrm{i}} \overline{a}^{\mathrm{ii}}_{x+t;z}.$$

Package of LTC and lifetime-related benefits

When $t < n$, the reserve at time t for an autonomous individual is equal to

$$V^{\mathrm{a}}_t = b_{\mathrm{a}} v(t,n)_{n-t} p^{\mathrm{aa}}_{x+t} \overline{a}^{\mathrm{aa}}_{x+n} + b_{\mathrm{i}} \overline{a}^{\mathrm{ai}}_{x+t} + c_{\mathrm{d}} \left(\overline{A}^{\mathrm{a;a\to d}}_{x+t} + \overline{A}^{\mathrm{a;i\to d}}_{x+t} \right) - \pi_{\mathrm{a}} \overline{a}^{\mathrm{aa}}_{x+t}.$$

The reserve for an individual in LTC state is equal to

$$V^{\mathrm{i}}_{t;z} = b_{\mathrm{i}} \overline{a}^{\mathrm{ii}}_{x+t;z} + c_{\mathrm{d}} \overline{A}^{\mathrm{i;i\to d}}_{x+t;z}.$$

6.11 Conclusion

In this chapter, we have explained how premiums and reserves for LTC insurance contracts can be computed. The equivalence principle inherited from life insurance remains at the heart of LTC insurance pricing. It has been applied here in a 3-state, Semi-Markov framework. Analytical expressions have been obtained for the premiums and reserves of different LTC products, including combined products. The impact of specific contract conditions on premiums and reserves has been quantified. For more details, we refer the

interested reader e.g. to Dickson et al. (2013), Haberman and Pitacco (1998) or Pitacco (2014); French-speaking readers may find convenient to refer to Denuit and Robert (2007).

Only annuity-type payouts have been considered here (a predetermined benefit level is assumed to be payable periodically to eligible individuals). This kind of product is typically sold in the EU. In some other countries, insurers sometimes reimburse the cost of assistance required because of the loss of autonomy rather than paying a predetermined monthly benefit. The 3-state model worked out in this chapter remains useful to forecast the likely start of payments and their duration but an additional model for claim severities is needed for pricing and reserving. In that respect, future trends in claim costs must be taken into account. Because medical inflation is typically hard to forecast, the management of such LTC insurance products becomes even more difficult.

This chapter goes beyond the basic LTC cover. We studied in particular the Belgian KBC product, which is very innovative in terms of pricing and risk management and could be considered as an inspiring example. The KBC package consists of a whole-life insurance coverage together with a LTC annuity with monthly payment m, subject to a deferred period. The length of the deferred period depends on policyholder's age x at policy issue. The product is sold as a combination of a life insurance contract and a LTC cover. Depending on the country, these two products may fall under different lines of business and must then be managed separately.

In this chapter, we have assumed that actual interest rates, morbidity and mortality rates remain equal to their assumed values entering actuarial formulas. In practice, these assumptions may be violated, sometimes to a large extent, and should be revised periodically in a dynamic perspective. We refer the reader to the indexing mechanism of Chapter 2 for contributions on this topic.

Very few studies investigated time trends in transition rates for multistate actuarial models. Renshaw and Haberman (2000) identified time trends using separate Poisson GLM regression models for each transition. Christiansen et al. (2012) and Levantesi and Menzietti (2012) allowed for possible correlations by means of a multivariate versions of mortality projection models (Lee-Carter and Cairns-Blake-Dowd models). Chapters 3 and 4 also both provide stochastic modelizations of morbidity trends with an age and time component.

The next chapter proposes an innovative P2P risk-sharing mechanism, enabling to mutualize longevity and morbidity risks.

References

- Brown, J. & Warshawsky, M. (2013). The life care annuity: A new empirical examination of an insurance innovation that addresses problems in the markets for life annuities and long-term care insurance. Journal of Risk and Insurance 80, 677-704.

- Christiansen, M., Denuit, M. & Lazar, D. (2012). The Solvency II square-root formula for systematic biometric risk. Insurance: Mathematics and Economics 50, 257-265.

- Denuit, M., Dhaene, J., Hanbali, H., Lucas, N. & Trufin, J. (2017). Updating mechanism for lifelong insurance contracts subject to medical inflation. European Actuarial Journal 7, 133-163.

- Denuit, M. & Robert, C. (2007). Actuariat des Assurances de Personnes: Modélisation, Tarification et Provisionnement. Collection Audit-Actuariat-Assurance, Economica, Paris.

- Dhaene, J., Godecharle, E., Antonio, K., Denuit, M. & Hanbali, H. (2017). Lifelong health insurance covers with surrender value: Updating mechanisms in the presence of medical inflation. ASTIN Bulletin 47, 803-836.

- Dickson, D., Hardy, M. & Waters, H. (2013). Actuarial Mathematics for Life Contingent Risks. Cambridge University Press.

- Haberman, S. & Pitacco, E (1998). Actuarial Models for Disability Insurance. CRC Press.

- Levantesi, S. & Menzietti, M. (2012). Managing longevity and disability risks in life annuities with long term care. Insurance: Mathematics and Economics 50, 391-401.

- Pitacco, E. (2014). Health Insurance: Basic Actuarial Models. Springer International Publishing.

- Pitacco, E. (2016). Premiums for Long-Term Care insurance packages: Sensitivity with respect to biometric assumptions. Risks 4, 1-22.

- Renshaw, A. & Haberman, S. (2000). Modelling the recent time trends in UK permanent health insurance recovery, mortality and claim inception transition intensities. Insurance: Mathematics and Economics 27, 365-396.

- Waters, H. (1990). The recursive calculation of the moments of the profit on a sickness insurance policy. Insurance: Mathematics and Economics 9, 101-113.

Chapter 7

Collaborative approach or P2P

The present chapter is based on the following paper:

- Hieber, P., Lucas, N. (2020). Life-care tontines (No. 2020026). UC Louvain, Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA). Submitted to Astin Bulletin on September 9, 2020.

7.1 Introduction

As we showed in the previous chapter, estimating the risks of a classical LTC cover or a life-care annuity is a challenging task for the insurance provider, resulting typically in high risk and administration charges. This might explain why the volume of the private market for LTC insurance is still relatively small. Indeed, when looking at the written gross premiums for long-term care insurance (LTCI), it is clear that the private LTC insurance market is limited in most OECD countries, although the need for a market is clearly strong (OECD (2020)).

In this chapter, we build on the advantage of pooling mortality and morbidity risk and we focus on mutual insurance, i.e. the risks are not taken by an insurance provider but shared within a pool of individuals. This significantly reduces charges but also leaves the risks to the pool members. A mutual insurance product would not guarantee a precise level of retirement income. On top of the investment returns from funded assets, survivors receive a higher payout funded by the "mortality credits" of deceased members. Classical mutual mortality risk pooling schemes are tontines (see, e.g., Sabin (2010), Forman & Sabin (2015), Milevsky & Salisbury (2015), Forman & Sabin (2016), Fullmer & Sabin (2018), Li & Rothschild (2019), Chen, Hieber & Rach (2020)) and pooled annuities (see, e.g., Piggott et al. (2005), Valdez et al. (2006), Stamos (2008), Qiao & Sherris (2013), Donnelly, Guillén & Nielsen (2013), Donnelly, Guillén & Nielsen (2014)). Instead of purely investing in a mutual insurance scheme, it might make sense to combine traditional retirement products and mutual insurance (see, e.g., Weinert & Gründl (2017),

Chen, Hieber & Klein (2019), Chen, Rach & Sehner (2020)). We introduce a "life-care tontine", which in addition to retirement income targets the needs of LTC coverage for an ageing population. The risk groups of a life-care tontine are not fixed but dynamic: people moving to dependency are assigned higher death probabilities, allowing them to get a bigger share in future mortality credits redistributed among the survivors of the tontine pool. To make the product attractive for subscribers with different risk, we suggest a fairness condition that ensures that the payments are actuarially fair *in each payment period* (see also Donnelly, Guillén & Nielsen (2013), Donnelly, Guillén & Nielsen (2014)). In other words, the life-care tontine stays fully funded at all times with each individual investment balance reflecting actual market values. We also allow to pool individuals from different age cohorts (see also Donnelly, Guillén & Nielsen (2014), Milevsky & Salisbury (2016), Denuit (2019)).

Such a product design has many advantages. (1) Compared to a life-care annuity, a life-care tontine has significantly lower solvency capital requirement (see also Shao et al. (2015), Chen, Hieber & Klein (2019)), inducing lower costs. (2) Compared to a classical tontine or pooled annuity, a life-care tontine is also attractive for people in poor health, reducing adverse selection costs. Further, according to Valdez et al. (2006), mutual insurance like a pooled annuity fund shows lower adverse selection relative to standard life annuities. (3) Being actuarially fair in each payment period, the life-care tontine avoids the disadvantage of a closed tontine pool (see, for example, the discussion in Chen, Hieber & Klein (2019)). The design allows to keep the pool size at a constant high level, replacing deceased individuals by new members. The sharing within the tontine pool is carried out by the concept of mortality and morbidity credits. (4) Pooling heterogeneous risks, i.e. different age-cohorts or active/dependent states, allows to increase tontine pool sizes and thus to reduce the overall risk. In our tontine scheme, individuals might change their risk classification, i.e. by moving from an "active" to a "dependent" state.

In Section 7.2, we introduce a 2-state alive/dead framework through a fair tontine scheme allowing members to freely join the pool. This framework enables to pool heterogeneous cohorts, like in Donnelly, Guillén & Nielsen (2014), Milevsky & Salisbury (2016) and Denuit (2019). Section 7.3 extends this to a 3-state framework, with a dependent state getting a specific (higher) payoff. The classical life-care annuity is compared with our life-care tontine. The fairness of the product is demonstrated and the payoffs are smoothed over time to fit the actual needs. Section 7.4 conclude and make additional remarks.

7.2 2-state framework

In a first step, we consider a 2-state framework where individuals have two possible states "alive" or "dead". Let us introduce the set of all individuals at initiation by $\mathscr{L}_0 = \{1,2,...,n\}$. Time is discretized in periods $t = 0,1,2,\ldots$. Assume that individual $j \in \mathscr{L}_0$, aged x_j with a remaining lifetime T_j, contributes a single premium $c_j(0)$ at time 0. In this article, we focus on mortality risk. Financial assets are invested in risk-free zero coupons and $v(s,t)$ is the discount factor from s to t. The maximal age is denoted by ω. For now,

the remaining lifetimes T_j, $j \in \mathscr{L}_0$, are assumed to be independent.

7.2.1 Tontine payoff

The n individuals form a tontine pool. Given the total initial premium payment, they decide on a withdrawal plan for the pool, that is for $t = 0,1,2,\ldots$, they (together) withdraw the amount $W_j(t)$ in a way that the premium equivalence

$$\underbrace{\sum_{j=1}^{n} c_j(0)}_{\text{total contributions}} = \sum_{j=1}^{n} \underbrace{\sum_{t=1}^{\omega-x_j} v(0,t)W_j(t)}_{\text{discounted benefits individual } j} \tag{7.2.1}$$

holds. The account value left according to the agreed decumulation plan for individual j at time $t = 0,1,2...$ is denoted $c_j(t)$. Equation (7.2.1) shows the main property of a tontine: the *sum* of all payoffs to the pool is deterministic, leaving no risk for the insurance provider. The payoff to a single individual $W_j(t)$, however, is random and may depend on the mortality experience in the pool. In the remainder of this section, we will demonstrate that (7.2.1) holds also at later points in time, that is the tontine scheme is fully funded at all times and satisfies:

$$\underbrace{\sum_{j=1}^{n} c_j(t)}_{\text{total account values}} = \sum_{j=1}^{n} \underbrace{\sum_{s=t+1}^{\omega-x_j} v(t,s)W_j(s)}_{\text{discounted future benefits individual } j} \tag{7.2.2}$$

We proceed by iteration to obtain $\mathscr{L}_t = \{j \in \mathscr{L}_0 \,|\, T_j > t\}$, the subset of participants still alive at time t. Let us define $\mathscr{D}_t = \{j \in \mathscr{L}_0 \,|\, t-1 < T_j \leq t\} = \mathscr{L}_{t-1} - \mathscr{L}_t$, the subset of participants dying in $(t-1,t]$. We denote by ${}_tp_{x_j} = \mathrm{E}[\mathbb{1}_{T_j>t}] = \mathrm{E}[\mathbb{1}_{j\in\mathscr{L}_t}]$ the probability for individual j to survive t years and set ${}_tq_{x_j} := 1 - {}_tp_{x_j}$. For annual survival and death probabilities, we abbreviate $p_{x_j} := {}_1p_{x_j}$ and $q_{x_j} := {}_1q_{x_j}$. For $t = 1,2,\ldots,\omega - x_j$, we obtain $\mathbb{1}_{j\in\mathscr{L}_t} \sim \mathrm{Ber}({}_tp_{x_j})$ and $\mathbb{1}_{j\in\mathscr{D}_t} \,|\, \{j \in \mathscr{L}_{t-1}\} \sim \mathrm{Ber}(q_{x_j+t-1})$. Note that our assumption of a maximal age ω implies that individuals never reach age $\omega+1$, that is $q_\omega = 1$.

Let us now look at an individual $j \in \mathscr{L}_{t-1}$ and a single time period $(t-1,t]$. During the time period $(t-1,t]$, the individual j's account value accrues to an amount of $v(t-1,t)c_j(t-1)$. In case of death in $(t-1,t]$, this account value is lost and distributed to the pool of individuals. Otherwise, the individual receives a payment at time t. This payment is decomposed into a fixed withdrawal $s_j(t)$ and mortality credits from deceased pool members. Each individual's account value is iteratively determined via

$$c_j(t) = \begin{cases} v(t-1,t)c_j(t-1) - s_j(t), & j \in \mathscr{L}_t \\ 0, & \text{otherwise} \end{cases} \tag{7.2.3}$$

in a way that the account value is depleted at the maximal age ω, that is $c_j(\omega - x_j) = 0$.

With this, we can solve (7.2.3) to get, for individual $j \in \mathscr{L}_t$ at time t:

$$c_j(t) = \sum_{u=t+1}^{\omega - x_j} v(t,u) s_j(u)\,. \tag{7.2.4}$$

To define the variable part of the payoff (the mortality credits), formally, denote as

$$X_j(t) := \mathbb{1}_{j \in \mathscr{D}_t} \cdot v(t-1,t)^{-1} c_j(t-1)$$

the random variable that is 0 in the case where the individual is alive at time t and equal to the accrued account value $v(t-1,t)^{-1}c_j(t-1)$ in case of death in $(t-1,t]$. At each time $t = 1,2,\ldots$, we have to distribute the pool's total mortality credit

$$X(t) := \sum_{j \in \mathscr{L}_{t-1}} X_j(t) = \sum_{j \in \mathscr{D}_t} v(t-1,t)^{-1} c_j(t-1)$$

among the individuals $j \in \mathscr{L}_{t-1}$ according to some predefined rule. In what follows $\beta_j(X(t))$ relates to a fair distribution rule. Its properties are defined later in this section.

The annual payoff to individual j is denoted by $W_j(t)$ (see above). At time t and for an individual $j \in \mathscr{L}_{t-1}$, it is given by:

$$W_j(t) = \begin{cases} s_j(t) + \beta_j(X(t)), & \text{if } j \in \mathscr{L}_t \\ \beta_j(X(t)), & \text{if } j \in \mathscr{D}_t \end{cases} \tag{7.2.5}$$

decomposed of

- $s_j(t)$: individual, fixed withdrawal amount,
- $\beta_j(X(t))$: collective part of the benefits, i.e. the mortality credits.

Note that the fixed withdrawal amount $s_j(t)$ is received only if the individual survives until time t. The individual always receives the mortality credit $\beta_j(X(t))$ – either to increase the fixed payoff (if $j \in \mathscr{L}_t$) or as a death benefit (if $j \in \mathscr{D}_t$). With (7.2.1), (7.2.4) and (7.2.5), it is possible to show that the scheme remains fully funded, i.e. the sum of individual account values at each time t is equal to the sum of discounted future benefits, see (7.2.2). In Definition 7.2.1, we define properties of a fair distribution rule $\beta_j(X(t))$, see also, for example, Denuit (2019). At the end of this section, we demonstrate how these properties lead to an actuarially fair tontine product.

Definition 7.2.1 (Fair distribution rule: mortality credits)**.** If the share distributed to individual $j \in \mathscr{L}_{t-1}$ is denoted by $\beta_j(X(t))$, a fair distribution rule has to satisfy the following properties:

- Self-sufficiency property: $\sum_{j \in \mathscr{L}_{t-1}} \beta_j(X(t)) = X(t)$.
- Positivity property: $\beta_j(X(t)) \geq 0$.

- Fairness property:

$$\mathrm{E}_{t-1}[\beta_j(X(t))] \quad = \quad \underbrace{\mathrm{E}_{t-1}[\mathbb{1}_{j\in\mathscr{D}_t}]}_{\text{probability to die in } (t-1,t]} \quad \cdot \underbrace{v(t-1,t)^{-1}c_j(t-1)}_{\text{amount at risk at time } t}, \tag{7.2.6}$$

where $\mathrm{E}_t := \mathrm{E}[\cdot\,|\,\mathscr{F}_t]$ is an expectation conditional on the information $\mathscr{F}_t := \sigma(\mathscr{L}_t)$.

In the 2-state framework, we have that $\mathrm{E}_{t-1}[\mathbb{1}_{j\in\mathscr{D}_t}] = q_{x_j+t-1}$, the probability that an individual is going to die in the time interval $(t-1,t]$. Fairness implies that on average he receives the same payoff whether he joins the tontine pool or not. In the first case, he receives $\beta_j(X(t))$, in the latter case $X_j(t)$, resulting in the fairness condition $\mathrm{E}_{t-1}[X_j(t)] = \mathrm{E}_{t-1}[\beta_j(X(t))]$, see (7.2.6). Thus, to be fair, on average, any individual $j \in \mathscr{L}_{t-1}$ receives the amount (7.2.6), which is on average proportional to both the death probability and the account value. Three examples of a fair distribution rule are presented in Examples 7.2.1–7.2.3, see also, e.g., Denuit & Robert (2020).

Example 7.2.1 (Conditional mean risk sharing rule)**.** At time t, each individual $j \in \mathscr{L}_{t-1}$ receives the mortality credit (respectively death benefit):

$$\beta_j(X(t)) = \mathrm{E}_{t-1}[X_j(t)\,|\,X(t)]\,. \tag{7.2.7}$$

(see, e.g., Denuit & Dhaene (2012), Denuit (2019))

Example 7.2.2 (Linear risk sharing rule)**.** At time t, each individual $j \in \mathscr{L}_{t-1}$ receives the mortality credit (respectively death benefit):

$$\beta_j(X(t)) = \frac{q_{x_j+t-1}\cdot c_j(t-1)}{\sum_{j\in\mathscr{L}_{t-1}} q_{x_j+t-1}\cdot c_j(t-1)}\cdot X(t)\,. \tag{7.2.8}$$

(see, e.g., Donnelly, Guillén & Nielsen (2013), Donnelly, Guillén & Nielsen (2014) and Schumacher (2018))

Example 7.2.3 (Linear regression rule)**.** At time t, each individual $j \in \mathscr{L}_{t-1}$ receives the mortality credit (respectively death benefit):

$$\beta_j(X(t)) = \mathrm{E}_{t-1}[X_j(t)] + \frac{\mathrm{Cov}_{t-1}[X_j(t),X(t)]}{\mathrm{Var}_{t-1}[X(t)]}(X(t) - \mathrm{E}_{t-1}[X(t)])\,. \tag{7.2.9}$$

For a motivation and comparison between the 3 distribution rules, we refer the interested reader to Denuit & Robert (2020).

The withdrawal plan (7.2.5) needs to be defined, i.e. one needs to know how to distribute the fixed withdrawals $s_j(t)$ over time. The only requirements we have are the premium equivalence (7.2.1) and the fairness of the distribution rule in Definition 7.2.1. Keeping this as general as possible, we assume that individual j pays the premium $c_j(0)$ to receive an *average* payoff of $b_j(t)$, for $t = 1,2,\ldots,\omega - x_j$. The individual might, for example, ask for an (on average) constant payoff $b_j(t) \equiv b_j = \mathrm{E}_{t-1}[W_j(t)\,|\,j \in \mathscr{L}_t]$ (see also Remark 7.2.4 for a discussion on the choice of $b_j(t)$).

Remark 7.2.4 (Choice of $b_j(t)$ and adverse selection)**.** Note that the individual payoffs $b_j(t)$ allow for a lot of flexibility in the tontine designs as the payoff is specific to each individual. If each individual may freely choose the average payoff $b_j(t)$, one should pay special care to adverse selection. For example depending on their personal health state, people will be incited to ask for a different payoff. In order to avoid adverse selection, it makes sense to choose $b_j(t) \equiv b(t)$ equal for everybody in the pool.
There might be reasons to choose this payoff to be increasing with time due to a higher liquidity need at old ages (see, e.g., Weinert & Gründl (2017)) or the fact that individuals are risk-averse with respect to mortality risk (see, e.g., Milevsky & Salisbury (2015), Chen, Hieber & Rach (2020)). An individual with logarithmic preferences optimally chooses a constant payoff $b_j(t) \equiv b(t)$.
To determine the fixed withdrawals over time, let us have a closer look at the expected payoff of a survivor $j \in \mathscr{L}_t$:

$$\begin{aligned} \mathrm{E}_{t-1}[W_j(t) \,|\, j \in \mathscr{L}_t] &= \mathrm{E}_{t-1}[\mathbb{1}_{j\in\mathscr{L}_t} \cdot s_j(t) + \mathbb{1}_{j\in\mathscr{L}_{t-1}} \cdot \beta_j(X(t)) \,|\, j \in \mathscr{L}_t] \\ &= s_j(t) + \mathrm{E}_{t-1}[\beta_j(X(t))] \\ &= s_j(t) + q_{x_j+t-1} v(t-1,t)^{-1} c_j(t-1)\,. \end{aligned} \tag{7.2.10}$$

Therefore, if survivors want to receive on average a payoff $b_j(t)$ at time t, ones needs to set

$$s_j(t) + q_{x_j+t-1} v(t-1,t)^{-1} c_j(t-1) = b_j(t)\,. \tag{7.2.11}$$

As the maximal age is ω, we can, for each individual j, iteratively solve the set of equations (7.2.11) backwards in time to obtain:

$$s_j(t) = \begin{cases} \frac{b_j(t)}{1+q_{\omega-1}}\,, & \text{for } t = \omega - x_j \\ \frac{b_j(t) - q_{x_j+t-1} \sum_{u=t+1}^{\omega-x_j} v(t,u) s_j(u)}{1+q_{x_j+t-1}}\,, & \text{for } t = \omega - x_j - 1, \omega - x_j - 2, \ldots, 1 \end{cases} \tag{7.2.12}$$

The big advantage of the decomposition into a fixed and a variable payoff by the backwards iteration (7.2.12) is the fact that it depends on quantities related to individual j only and is independent of the other individuals in the pool. For a constant average payoff $b_j(t) \equiv b_j$, one typically obtains mortality credits that are increasing over time while the fixed payoff $s_j(t)$ is decreasing over time (see the numerical example in Section 7.2.3).

7.2.2 Actuarial fairness

Equations (7.2.5) and (7.2.12), together with one of the sharing rules from Examples 7.2.1-7.2.3, fully define the payoff of a tontine in a 2-state framework. The first advantage of this scheme is that it allows to pool policyholders with different mortality risks, for example from different age cohorts. The second advantage is that it is actuarially fair in each period: at each time t, the expected discounted future payoffs to any individual j equal this individual's current account value $c_j(t)$, see Theorem 7.1.

Theorem 7.1 (Actuarial fairness 2-state framework)**.** *The fairness condition* (7.2.6) *implies that the current account value* (7.2.3) *is actuarially fair at each time* $t = 0, 1, \ldots, \omega - x_j$*, that is:*

$$c_j(t) = \mathrm{E}_t\left[\sum_{k=t+1}^{\omega-x_j} v(t,k)W_j(k)\right]. \tag{7.2.13}$$

The conditional mean risk-sharing rule (7.2.7)*, the linear sharing rule* (7.2.8) *and the linear regression rule* (7.2.9) *satisfy the fairness condition* (7.2.6)*.*

Proof: At time $t = \omega - x_j$, individual j reaches the maximum possible age. The last year of life the individual only receives death benefits, and with (7.2.4) we get $c_j(\omega - x_j) = 0$. It implies that $c_j(\omega - x_j - 1) = v(\omega - x_j - 1, \omega - x_j)s_j(\omega - x_j)$.
We prove (7.2.13) by backwards induction. Assume that (7.2.13) holds for t. Using (7.2.3), (7.2.5) and (7.2.6), we find for an individual $j \in \mathscr{L}_{t-1}$ that:

$$\begin{aligned}
&\mathrm{E}_{t-1}\left[\sum_{k=t}^{\omega-x_j} v(t-1,k)W_j(k)\right] \\
&\quad = v(t-1,t)\Big(\mathrm{E}_{t-1}[W_j(t) + \mathbb{1}_{j\in\mathscr{L}_t}\cdot c_j(t)]\Big) \\
&\quad = v(t-1,t)\Big(\mathrm{E}_{t-1}[\mathbb{1}_{j\in\mathscr{L}_t}\cdot s_j(t) + \beta_j(X(t))] + p_{x_j+t-1}\cdot c_j(t)\Big) \\
&\quad = v(t-1,t)\Big(p_{x_j+t-1}\cdot s_j(t) + \mathrm{E}_{t-1}[\beta_j(X(t))] + p_{x_j+t-1}\cdot c_j(t)\Big) \\
&\quad = v(t-1,t)\Big(p_{x_j+t-1}\cdot s_j(t) + q_{x_j+t-1}\cdot v(t-1,t)^{-1}c_j(t-1) + p_{x_j+t-1}\cdot c_j(t)\Big) \\
&\quad = c_j(t-1)\,.
\end{aligned}$$

This shows that (7.2.13) also holds for $t-1$.
Condition (7.2.6) is satisfied for the conditional mean risk-sharing rule as for each individual $j \in \mathscr{L}_{t-1}$:

$$\begin{aligned}
\mathrm{E}_{t-1}[\beta_j(X(t))] &= \mathrm{E}_{t-1}\Big[\mathrm{E}_{t-1}[X_j(t)\,|\,X(t)]\Big] \\
&= \mathrm{E}_{t-1}[X_j(t)] = q_{x_j+t-1}\cdot v(t-1,t)^{-1}c_j(t-1)\,,
\end{aligned}$$

as well as for the linear risk-sharing rule as:

$$\begin{aligned}
\mathrm{E}_{t-1}[\beta_j(X(t))] &= \mathrm{E}_{t-1}\left[\frac{q_{x_j+t-1}\cdot c_j(t-1)}{\sum_{j=1}^n \mathbb{1}_{j\in\mathscr{L}_{t-1}}\cdot q_{x_j+t-1}\cdot c_j(t-1)}X(t)\right] \\
&= \frac{q_{x_j+t-1}\cdot c_j(t-1)}{\sum_{j\in\mathscr{L}_{t-1}} q_{x_j+t-1}\cdot c_j(t-1)}\cdot \mathrm{E}_{t-1}[X(t)] \\
&= q_{x_j+t-1}\cdot v(t-1,t)^{-1}c_j(t-1)\,,
\end{aligned}$$

and the linear regression rule:

$$\begin{aligned}\mathrm{E}_{t-1}[\beta_j(X(t))] &= \mathrm{E}_{t-1}\left[\mathrm{E}_{t-1}[X_j(t)] + \frac{\mathrm{Cov}_{t-1}[X_j(t),X(t)]}{\mathrm{Var}_{t-1}[X(t)]}(X(t)-\mathrm{E}_{t-1}[X(t)]\right] \\ &= \mathrm{E}_{t-1}[X_j(t)] = q_{x_j+t-1}\cdot v(t-1,t)^{-1}c_j(t-1)\,.\end{aligned}$$

□

Theorem 7.1 demonstrates that our tontine scheme allows to share mortality risk between heterogeneous individuals (i.e. individuals with different life expectancies), see also Donnelly, Guillén & Nielsen (2014), Milevsky & Salisbury (2015), Denuit (2019). The fact that the scheme is fair at each time point t gives a second advantage: the design allows individuals to later join the tontine scheme at an actuarially fair price. By design, joining the scheme does not affect the average benefits of the existing members. In contrast, in a closed tontine scheme, the number of pool members is decreasing over time, leading to an increase in risk at old ages (see, e.g., Chen, Hieber & Klein (2019)).

7.2.3 Numerical example 1

Let us illustrate our payoff in a numerical example, considering a pool of size $n = 10000$ where half of the pool has initial age 65 and half of the pool has initial age 85. For illustrative purposes, we choose the interest rate as $\delta_j = 0$ and an average payoff of $b_j(t) \equiv b_j = 1$ for both cohorts. The data correspond to values in line with observations made on the French LTC market. We apply the backward iteration (7.2.12) to obtain the fixed part of the payoff $s_j(t)$ and use (7.2.4) to get the account value $c_j(t)$ for $t = 1,2,\ldots,\omega - x_j$. Figure 7.1 gives the total payoff $W_j(t)$ and the fixed part of the payoff $s_j(t)$ for an individual from the 65-year cohort (left) and the 85-year cohort (right). For the payoff $W_j(t)$, we plot one random path. We observe that mortality credits are increasing over time and are higher for the 85-year cohort. Based on 10 000 simulations, averages and the 95% confidence interval are obtained on Figures 7.2 and 7.3. Figure 7.4 shows the individual account value $c_j(t)$ for both cohorts. According to Theorem 7.1, this account value is equal to the expected discounted value of future payoffs for individual j.

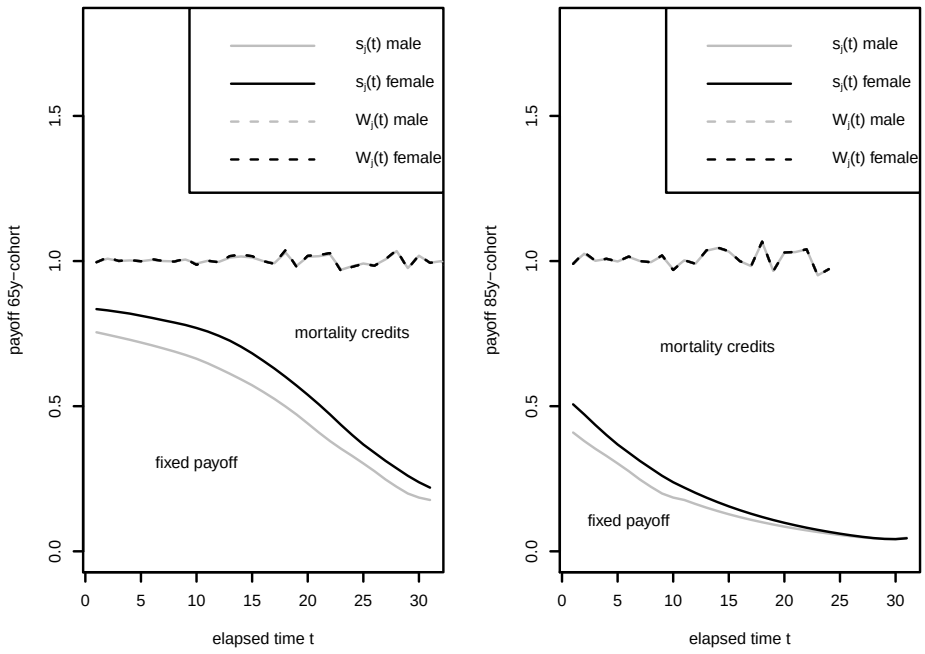

Figure 7.1: Evolution of fixed withdrawal $s_j(t)$ and total payoff $W_j(t)$ (one simulation path), young (left) and old cohort (right). We use the conditional mean risk sharing rule for this illustrative example.

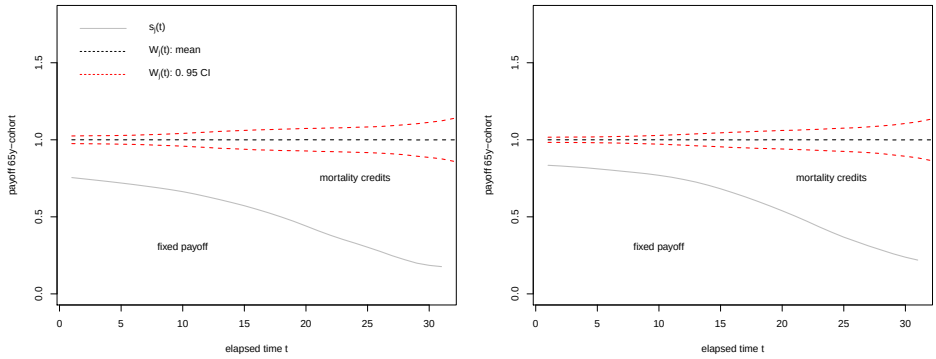

Figure 7.2: Young cohort: evolution of fixed withdrawal $s_j(t)$ and average total payoff $E[W_j(t)]$ (10 000 simulation paths) with 95% CI, male (left) and female (right).

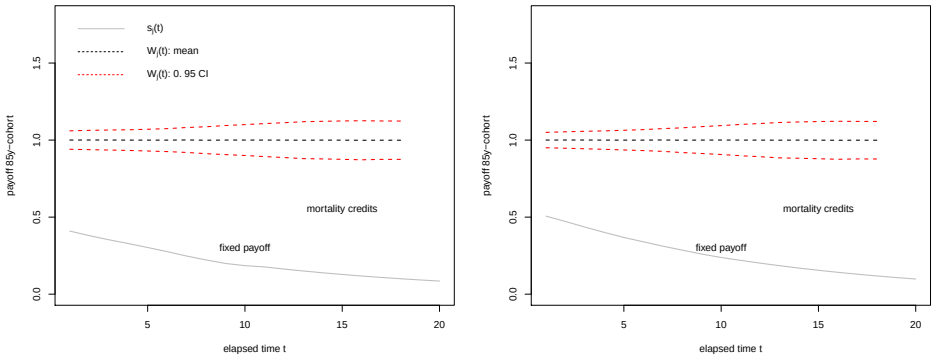

Figure 7.3: Old cohort: evolution of fixed withdrawal $s_j(t)$ and average total payoff $E[W_j(t)]$ (10 000 simulation paths) with 95% CI, male (left) and female (right).

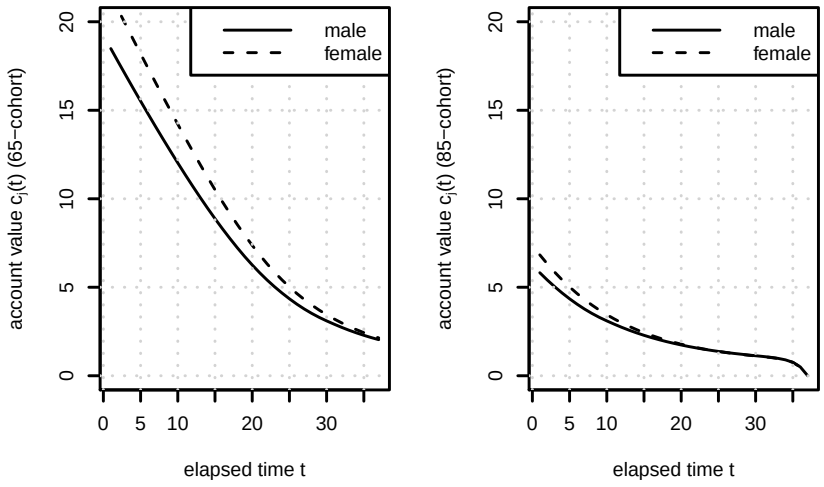

Figure 7.4: Evolution of the personal account $c_j(t)$ with time, young (left) and old cohort (right).

7.3 3-state framework

In a second step, we consider a 3-state semi-Markov model where any individual is either active (a), dependent (i) or dead (d). Initially, each individual is assumed to be in state active. In Section 7.3.1, we introduce additional notation for the 3-state model. We discuss the payoff of a life-care annuity in Section 7.3.2 before introducing our life-care tontine product in Section 7.3.3.

7.3.1 Additional notation

For an x_j-year old individual, let us define:

1. ${}_tp_{x_j}^{aa}$: the t-period sojourn probability in active state.

2. ${}_tp_{x_j}^{ai}$: the t-period transition probability from state a to i.

3. ${}_1p_{x_j}^{ad} = q_{x_j}^{(a)}$ and ${}_1p_{x_j;z}^{id} = q_{x_j;z}^{(i)}$: the annual death probabilities in state a and i, respectively. It is semi-Markovian in the latter case, with $z = 0,1,2\ldots$ the time already spent in dependency.

The individual's remaining lifetime T_j is decomposed into:

$$T_j = T_j^{(a)} + T_j^{(i)}, \tag{7.3.1}$$

where $T_j^{(a)}$ is the time spent in autonomy and $T_j^{(i)}$ is the time spent in dependence or disability. We have:

$$P(T_j^{(i)} = 0) > 0. \tag{7.3.2}$$

Let us define the number of individuals in the active and dependent state, respectively, at a future time t:

$$\mathscr{A}_t := \{j \in \mathscr{L}_t \,|\, T_j^{(a)} > t\}, \tag{7.3.3}$$

$$\mathscr{I}_{t;z} := \{j \in \mathscr{L}_t \,|\, T_j^{(a)} \leq t, T_j > t, z = t - T_j^{(a)}\}, \tag{7.3.4}$$

$$\mathscr{I}_t := \cup_{z=0}^{t-1} \mathscr{I}_{t;z} = \{j \in \mathscr{L}_t \,|\, T_j^{(a)} \leq t, T_j > t\} = \mathscr{L}_t \setminus \mathscr{A}_t. \tag{7.3.5}$$

Relating this to the notation above, this means that ${}_tp_{x_j}^{aa} = \mathrm{E}[\mathbb{1}_{j\in\mathscr{A}_t}]$, ${}_tp_{x_j}^{ai} = \mathrm{E}[\mathbb{1}_{j\in\mathscr{I}_t}]$, $q_{x_j+t-1}^{(a)} = \mathrm{E}[\mathbb{1}_{j\in\mathscr{D}_t\cup\mathscr{A}_{t-1}}]$, $q_{x_j+t-1;z}^{(i)} = \mathrm{E}[\mathbb{1}_{j\in\mathscr{D}_t\cup\mathscr{I}_{t-1;z-1}}]$, and $p_{x_j+t-1;z-1}^{ii} = \mathrm{E}[\mathbb{1}_{j\in\mathscr{L}_t\cup\mathscr{I}_{t-1;z-1}}]$.

7.3.2 Life-care annuity

In this section, we introduce life-care annuities and base ourselves on the works of, for example, Murtaugh et al. (2001), Spillman et al. (2003), Rickayzen (2007), Brown & Warshawsky (2013), Shao et al. (2015) and Chen et al. (2020). In contrast to the mutual insurance scheme discussed in this article, in a life-care annuity, mortality and morbidity risks are taken by an insurance provider. Each individual j pays the single premium $c_j(0)$ to buy an annuity with a future payment stream of $b_j(t), t = 1,2,\ldots,\omega - x_j$.

This annuity is supplemented with an LTC cover that provides an annual amount of $(\alpha_j - 1)\cdot b_j(t)$ as long as people are dependent. $\alpha_j > 1$ is an individual-specific constant reflecting an increased payoff in dependency. This additional LTC cover is an LTC annuity where the risk is taken by the insurance company. Ignoring administration and risk charges, the fair single premium $c_j(0)$ of the life-care annuity is given by:

$$c_j(0) = \sum_{t=1}^{\omega - x_j} ({}_tp_{x_j}^{ai}\, v(0,t)\alpha_j \cdot b_j(t) + {}_tp_{x_j}^{aa}\, v(0,t) b_j(t)) \,. \tag{7.3.6}$$

7.3.3 Life-care tontine

Based on the tontine scheme introduced in Section 7.2, we presents a life-care tontine that on average provides the same payout as the life-care annuity from the previous Section 7.3.2. In a life-care tontine, payments are adapted according to the autonomy/dependence of an individual. We define by $c_j^{(a)}(t)$ and $c_j^{(i)}(t;z)$ the current account values of an active and dependent individual, respectively. Assuming that, at time 0, every individual is autonomous, we set $c_j^{(a)}(0) = c_j(0)$. The main idea is that individuals moving into the dependent state have a higher death probability than people staying in active state. If mortality credits in a tontine scheme account for this increase, the payments in dependency naturally increase. To define payments in a life-care tontine for an individual $j \in \mathscr{L}_{t-1}$, we modify the fairness condition (7.2.6) to distinguish between active ($j \in \mathscr{A}_{t-1}$) and dependent individuals ($j \in \mathscr{I}_{t-1;z}$), with z the time spent in dependency (in years):

$$\mathrm{E}_{t-1}[\beta_j(X(t)) \,|\, j \in \mathscr{A}_{t-1}] = q_{x_j+t-1}^{(a)} \cdot v(t-1,t)^{-1} c_j^{(a)}(t-1) \,, \tag{7.3.7}$$

$$\mathrm{E}_{t-1}[\beta_j(X(t)) \,|\, j \in \mathscr{I}_{t-1;z}] = q_{x_j+t-1;z}^{(i)} \cdot v(t-1,t)^{-1} c_j^{(i)}(t-1;z) \,, \tag{7.3.8}$$

where, from now on, $\mathrm{E}_t := \mathrm{E}[\cdot \,|\, \mathscr{F}_t]$ is an expectation conditional on the information $\mathscr{F}_t := \sigma(\mathscr{A}_t, \mathscr{I}_{t;0}, \mathscr{I}_{t;1}, \ldots, \mathscr{I}_{t;t-1})$. With this design, we apply Definition 7.2.1 to the 3-state framework. The increased death probability in dependency ($q_{x_j+t-1;z}^{(i)} > q_{x_j+t-1}^{(a)}$) increases the share of mortality credits and thus the overall payoff as soon as an individual moves from the active to the dependent state.

Again, the cash-flows satisfy the premium equivalence (7.2.1). In a tontine, the payoff to the pool (left hand side of (7.2.1)) is fixed, leaving the insurance provider with no mortality nor morbidity risk. The payoffs to the pool members $W_j(t)$ are random and depend on the mortality and morbidity in the pool.

Adjusting mortality credits to dependency

Mortality credits are now distributed according to the individual's state (active, dependent, dead) using the fairness condition (7.3.7) and (7.3.8). We aim for an average payoff $\alpha_j(T^{(a)}) \cdot b_j(t)$ in dependency, where in this chapter $\alpha_j(T^{(a)})$ is a constant that depends on the time spent in the active state. In our notation, this means that:

$$\mathrm{E}[W_j(t) \,|\, j \in \mathscr{A}_t] = b_j(t)\,, \tag{7.3.9}$$

$$\mathrm{E}[W_j(t) \,|\, j \in \mathscr{I}_{t;t-T^{(a)}}] = \alpha_j(T^{(a)}) \cdot b_j(t)\,, \quad t \geq T^{(a)}. \tag{7.3.10}$$

To achieve the desired average payoff (7.3.9) and (7.3.10) in the active and dependent state, respectively, we – as in Section 7.2 – decompose the payoff in a fixed and a variable part. The fixed part of individual j in the active and dependent state is denoted by $s_j^{(a)}(t)$ and $s_j^{(i)}(t;z)$, respectively. The pool observes time-t withdrawals $W_j(t)$. For an individual $j \in \mathscr{L}_{t-1}$:

$$W_j(t) = \begin{cases} s_j^{(a)}(t) + \beta_j(X(t))\,, & \text{if } j \in \mathscr{A}_t \\ s_j^{(i)}(t;z) + \beta_j(X(t))\,, & \text{if } j \in \mathscr{I}_{t;z} \\ \beta_j(X(t))\,, & \text{if } j \in \mathscr{D}_t \end{cases} \tag{7.3.11}$$

Starting with an initial account value of $c_j^{(a)}(0) = c_j(0)$, the account for an active individual $j \in \mathscr{A}_{t-1}$ ($t \leq T^{(a)}, z \geq 1$) evolves as in the 2-state framework, see (7.2.3):

$$c_j^{(a)}(t) = \begin{cases} v(t-1,t)^{-1} c_j^{(a)}(t-1) - s_j^{(a)}(t)\,, & j \in \mathscr{A}_t \text{ and } t < T^{(a)} \\ v(t-1,t)^{-1} c_j^{(a)}(t-1) - s_j^{(i)}(t;0)\,, & j \in \mathscr{I}_{t;0} \text{ and } t = T^{(a)} \\ 0\,, & \text{otherwise} \end{cases} \tag{7.3.12}$$

The state-dependent constant $\alpha_j(T^{(a)})$ is chosen in a way that the product is actuarially fair, that is, at the time $T^{(a)}$ that an individual moves into dependency, the account value does not change:

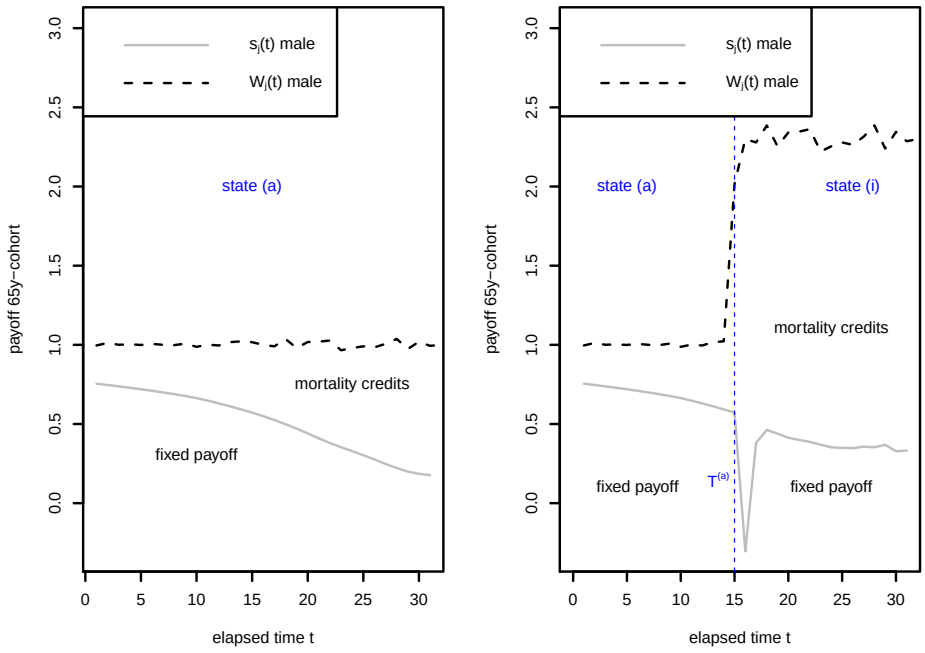

Figure 7.5: Evolution of fixed withdrawal $s_j^{(a)}(t)$ and $s_j^{(i)}(t;t-T^{(a)})$ and total payoff $W_j(t)$ (one simulation path), $x_j = 65, T^{(a)} = \omega - x_j$ (left) and $T^{(a)} = 15$ (right).

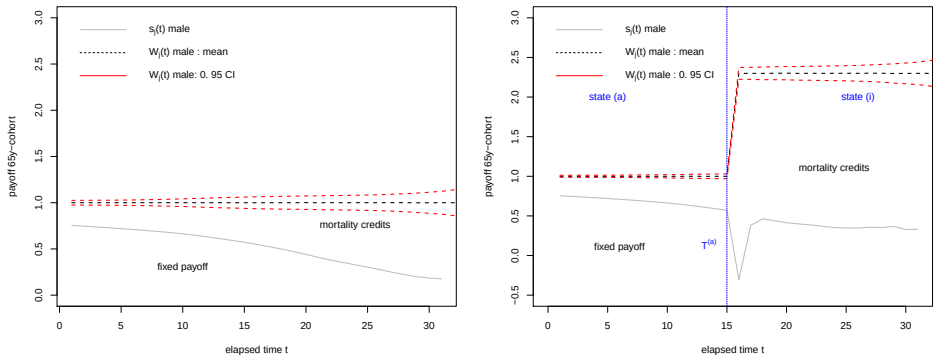

Figure 7.6: Evolution of fixed withdrawal $s_j^{(a)}(t)$ and $s_j^{(i)}(t;t-T^{(a)})$ and average total payoff $E[W_j(t)]$ (10 000 simulation paths) with 95% CI, $x_j = 65, T^{(a)} = \omega - x_j$ (left) and $T^{(a)} = 15$ (right).

$$
\begin{aligned}
&c_j^{(i)}(T^{(a)};0) + (\alpha_j(T^{(a)}) - 1) b_j(T^{(a)}) \\
&\qquad = \mathrm{E}_{T^{(a)}}\left[\sum_{k=T^{(a)}+1}^{\omega - x_j} v(T^{(a)},k) W_j(k) \,\middle|\, j \in \mathscr{I}_{T^{(a)};0} \right] + (\alpha_j(T^{(a)}) - 1) b_j(T^{(a)}) \\
&\qquad = \mathrm{E}_{T^{(a)}}\left[\sum_{k=T^{(a)}+1}^{\omega - x_j} v(T^{(a)},k) W_j(k) \,\middle|\, j \in \mathscr{A}_{T^{(a)}} \right] = c_j^{(a)}(T^{(a)}) \,. \qquad (7.3.13)
\end{aligned}
$$

We choose the constants $\alpha_j(T^{(a)})$ such that (7.3.13) is satisfied. In dependency ($t > T^{(a)}$, $j \in \mathscr{I}_{t-1}$), the account value evolves as follows:

$$
c_j^{(i)}(t;z) = \begin{cases} v(t-1,t)^{-1} c_j^{(i)}(t-1;z-1) - s_j^{(i)}(t;z) \,, & j \in \mathscr{I}_t \\ 0 \,, & \text{otherwise} \end{cases} \qquad (7.3.14)
$$

The way to determine the payoff decomposition is presented in Theorem 7.2. Figure 7.5 gives a sample path for an active male person with an average payoff of $b_j(t) = 1$ (left) and an individual that moves into dependency at time $T^{(a)} = 15$ (right) and Figure 7.6 shows the average payoff together with a 95% confidence interval based on 10000 simulations. The first years after moving into dependency are typically accompanied by a strong increase in mortality. In this case, the fixed part of the payoff even turns negative.

Theorem 7.2 (Choice of $\alpha_j(T^{(a)})$, $s_j^{(a)}(t)$, $s_j^{(i)}(t;t-T^{(a)})$)**.**
Consider an annual time grid $t \in \mathbb{N}$. An active individual ($j \in \mathscr{A}_t$) receives the fixed payoff

$s_j^{(a)}(t)$ *determined via the backwards iteration:*

$$s_j^{(a)}(t) = \begin{cases} \frac{b_j(t)}{1+q^{(a)}_{\omega-1}}, & \text{for } t = \omega - x_j \\ \frac{b_j(t) - q^{(a)}_{x_j+t-1} \sum_{u=t+1}^{\omega-x_j} v(t,u) s_j^{(a)}(u)}{1+q^{(a)}_{x_j+t-1}}, & \text{for } 1 \le t < \omega - x_j \end{cases} \tag{7.3.15}$$

A dependent individual that spent $t - T^{(a)}$ *years in dependency (*$j \in \mathscr{I}_{t;t-T^{(a)}}$*), receives for time* $t \ge T^{(a)}$ *the fixed payoff*

$$s_j^{(i)}(t; t - T^{(a)}) = \alpha_j(T^{(a)}) \cdot \widetilde{s}_j^{(i)}(t; t - T^{(a)}), \tag{7.3.16}$$

where $\widetilde{s}_j^{(i)}(t; t - T^{(a)})$ *is, for* $t \ge T^{(a)}$*, determined via the backwards iteration:*

$$\widetilde{s}_j^{(i)}(t; t - T^{(a)}) = \begin{cases} \frac{b_j(t)}{1+q^{(i)}_{\omega-1;t-T^{(a)}-1}}, & \text{for } t = \omega - x_j \\ \frac{b_j(t) - q^{(i)}_{x_j+t-1;t-T^{(a)}-1} \sum_{u=t+1}^{\omega-x_j} v(t,u) \widetilde{s}_j^{(i)}(u; u-T^{(a)})}{1+q^{(i)}_{x_j+t-1;t-T^{(a)}-1}}, & \text{for } T^{(a)} \le t < \omega - x_j \end{cases} \tag{7.3.17}$$

The factor $\alpha_j(T^{(a)})$ *that increases payments in dependency is determined via:*

$$\alpha_j(T^{(a)}) = \frac{\sum_{u=t+1}^{\omega-x_j} v(t,u) s_j^{(a)}(u) + b_j(t)}{\sum_{u=t+1}^{\omega-x_j} v(t,u) \widetilde{s}_j^{(i)}(u; u - T^{(a)}) + b_j(t)}. \tag{7.3.18}$$

Proof: The payoff $\widetilde{s}_j^{(i)}(t; t - T^{(a)})$ for a dependent individual $j \in \mathscr{I}_t$ receiving an average payoff $b_j(t)$ at times $t \ge T^{(a)}$ is obtained using the 2-state semi-Markov backwards iteration system (7.3.15), see also the similar iteration in Section 7.2, Equation (7.2.12). As we do not allow for additional payments in dependency, we want to choose $\alpha_j(T^{(a)})$ in (7.3.16) such that the present value of future payoffs does not change if a person moves to dependency, i.e. (7.3.13) is satisfied. This implies, for an active individual $j \in A_{t-1}$:

$$\begin{aligned} c_j^{(a)}(t) &= \begin{cases} v(t-1,t)^{-1} c_j^{(a)}(t-1) - s_j^{(a)}(t), & \text{if } j \in \mathscr{A}_t \\ c_j^{(i)}(t;0), & \text{if } j \in \mathscr{I}_t \\ 0, & \text{if } j \in \mathscr{D}_t \end{cases} \\ &= \begin{cases} v(t-1,t)^{-1} c_j^{(a)}(t-1) - s_j^{(a)}(t), & \text{if } j \in \mathscr{L}_t \\ 0, & \text{if } j \in \mathscr{D}_t \end{cases} \end{aligned}$$

As in Section 7.2 Equation (7.2.12), we can solve this system to obtain:

$$c_j^{(a)}(t) = \sum_{u=t+1}^{T} v(t,u)\, s_j^{(a)}(u)\,. \tag{7.3.19}$$

The backward iteration (7.3.15) determines the fixed part of the payoff $s_j^{(a)}(t)$ for an active individual, see also the 2-state framework in Section 7.2, Equation (7.2.12).
Let us name $\widetilde{c}_j^{(i)}(t;t-T^{(a)})$ the reference amount, based on a predetermined $\alpha_j(T^{(a)})$-value of 1 and the corresponding fixed payments $\widetilde{s}_j^{(i)}(t;t-T^{(a)})$. We have

$$s_j^{(i)}(t;t-T^{(a)}) = \alpha_j(T^{(a)}) \cdot \widetilde{s}_j^{(i)}(t;t-T^{(a)})\,.$$

It is deduced that

$$c_j^{(i)}(t;t-T^{(a)}) = \alpha_j(T^{(a)}) \cdot \widetilde{c}_j^{(i)}(t;t-T^{(a)})\,.$$

We solve for α in (7.3.13). If we use (7.3.13), that is if we assume that the present value of future payoffs is unchanged if a person moves into dependency, we obtain

$$\alpha_j(T^{(a)}) = \frac{c^{(a)}(T^{(a)}) + b_j(t)}{\widetilde{c}^{(i)}(T^{(a)};0) + b_j(t)} = \frac{\sum_{u=t+1}^{\omega-x_j} v(t,u) s_j^{(a)}(u) + b_j(t)}{\sum_{u=t+1}^{\omega-x_j} v(t,u)\, \widetilde{s}_j^{(i)}(u;u-T^{(a)}) + b_j(t)}\,.$$

□

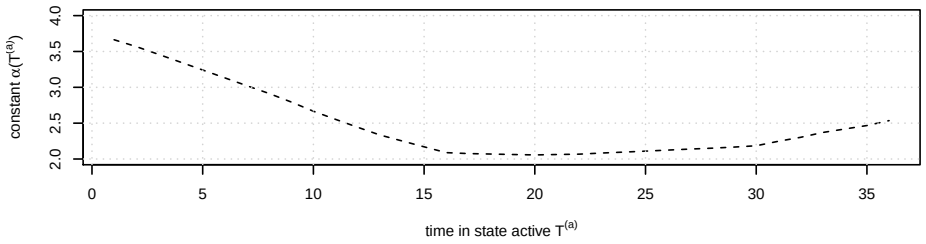

Figure 7.7: Adjustment constant $\alpha_j(T^{(a)})$ as a function of the time in the active state $T^{(a)}$ (if $T^{(i)} > 0$).

Figure 7.7 presents the function $\alpha_j(T^{(a)})$ in our data set. The data correspond to values in line with observations made on the French LTC market. If $\alpha_j(T^{(a)}) = 1$, this would mean

that an individual in dependency would receive, on average, the same payoff as if he/she were active.

We observe that $\alpha_j(T^{(a)})$ takes values between 2 and 4 which implies a considerable increase of benefits in dependency, that is a dependent individual may receive a 2-4 times higher payoff than an active individual. The increase strongly depends on the time $T^{(a)}$ the person moves into dependency. If we want to fix the increase in dependency, say to $\alpha_j(T^{(a)}) = \alpha_j$ as in the case of the life-care annuity in Section 7.3.2, we need to share the corresponding loss / gain that appears if somebody moves into dependency, see the following section.

A priori fixation of $\alpha_j(T^{(a)})$

As a next step, we want to fix the payoff in dependency with a predetermined increase in the dependent state to α_j. In other words, we want to smooth $\alpha_j(T^{(a)})$ from the previous section (see Figure 7.7). Formally, denote as

$$Y_j(t) := \mathbb{1}_{j\in\mathscr{I}_{t;0}}\Big((c_j^{(a)}(t) - c_j^{(i)}(t;0)) + (1-\alpha_j)b_j(t)\Big)$$

the morbidity credits for individual j. Morbidity credits are needed to adjust the benefits of individuals that have moved to the dependent state in $(t-1,t]$ and are still alive at time t (that is an individual $j \in \mathscr{I}_{t;0}$). They contain two parts: $(1-\alpha_j)b_j(t)$ increases the payoff at the first payoff date after moving into dependency while $(c_j^{(a)}(t) - c_j^{(i)}(t;0))$ adjusts the later payoffs. The morbidity credits are redistributed among the pool of individuals. Note that they can be positive or negative, depending on whether the $\alpha_j(T^{(a)})$-value is higher or lower than the "fair" increase determined in the previous section (see the values presented in Figure 7.7). At each time $t = 1,2,\ldots,T$, we have to distribute

$$Y(t) := \sum_{j\in\mathscr{A}_{t-1}} Y_j(t)$$

according to some predefined rule. We, similarly to the concept of mortality credits in the previous section, introduce a function $\gamma_j(Y(t))$ that redistributes the morbidity credits $Y(t)$ within the pool, see Definition 7.3.1.

Definition 7.3.1 (Fair distribution rule: morbidity credits)**.** If the share distributed to individual $j \in \mathscr{L}_{t-1}$ is denoted by $\gamma_j(Y(t))$, a fair distribution rule has to satisfy the following properties:

- Self-sufficiency property: $\sum_{j\in\mathscr{L}_{t-1}} \gamma_j(Y(t)) = Y(t)$.

- Fairness property:

$$\mathrm{E}_{t-1}[\gamma_j(Y(t))] \quad = \quad \underbrace{\mathrm{E}_{t-1}[\mathbb{1}_{j\in\mathscr{I}_{t;0}}]}_{\text{probability to get dependent in } (t-1,t]} \cdot \underbrace{(c_j^{(a)}(t) - c_j^{(i)}(t) + (1-\alpha_j)b_j(t))}_{\text{required capital at time } t}. \tag{7.3.20}$$

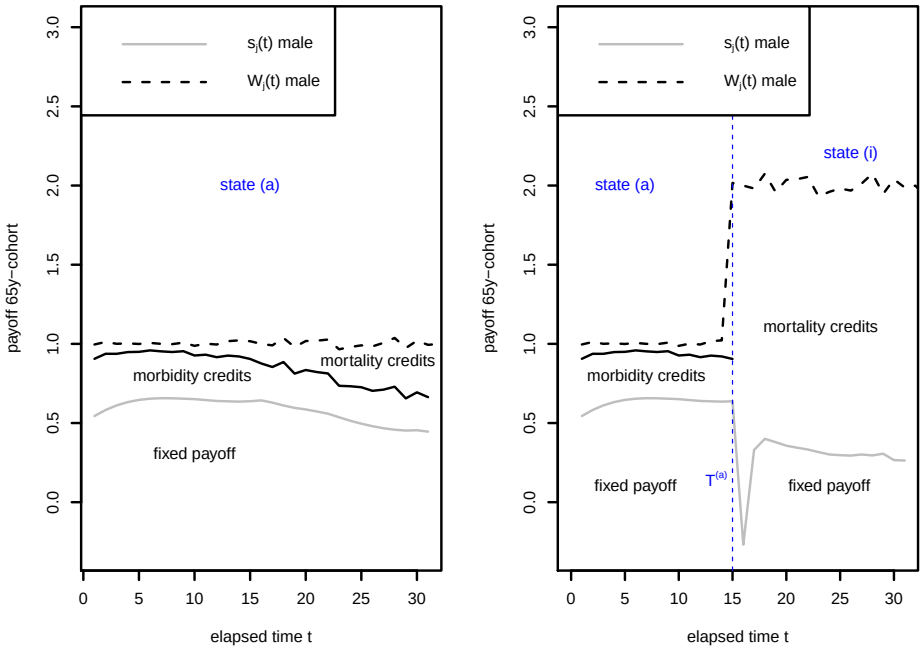

Figure 7.8: Evolution of fixed withdrawal $s_j^{(a)}(t)$ and $s_j^{(i)}(t;t-T^{(a)})$ and total payoff $W_j(t)$ (one simulation path), $x_j = 65, T^{(a)} = \omega - x_j$ (left) and $T^{(a)} = 15$ (right).

Again, we can, for example, choose a conditional mean risk-sharing, linear sharing or linear regression rule as a distribution rule $\gamma_j(\cdot)$. For an active individual, we can rewrite (7.3.20) to obtain

$$\mathrm{E}_{t-1}[\gamma_j(Y(t)) \,|\, j \in \mathscr{A}_{t-1}] \quad = p_{x_j+t-1}^{ai} \cdot (c_j^{(a)}(t) - c_j^{(i)}(t) + s_j^{(a)}(t) - s_j^{(i)}(t)). \tag{7.3.21}$$

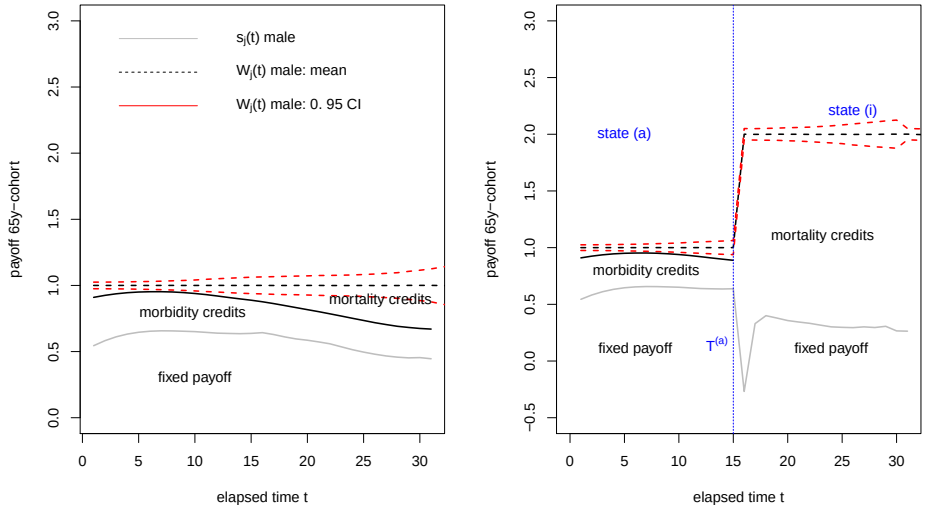

Figure 7.9: Evolution of fixed withdrawal $s_j^{(a)}(t)$ and $s_j^{(i)}(t;t-T^{(a)})$ and average total payoff $E[W_j(t)]$ (10 000 simulation path) with 95% CI, $x_j = 65, T^{(a)} = \omega - x_j$ (left) and $T^{(a)} = 15$ (right).

If the individual is dependent or dead already at time $t-1$, we obtain $\mathrm{E}_{t-1}[\gamma_j(Y(t)) \,|\, j \in \mathscr{I}_{t-1}] = \mathrm{E}_{t-1}[\gamma_j(Y(t)) \,|\, j \in \mathscr{D}_{t-1}] = 0$, that is in a fair distribution scheme dead or dependent people do (on average) not receive any morbidity credits. In our tontine scheme, we thus redistribute the credits among active individuals $j \in \mathscr{A}_{t-1}$ only. In a later extension, it might make sense to share the risk $Y(t) - \mathrm{E}_{t-1}[Y(t)]$ among all survivors $j \in \mathscr{L}_{t-1}$. The pool observes time-t withdrawals $W_j(t)$, decomposed into a fixed withdrawal, mortality and morbidity credits. For an active individual $j \in \mathscr{A}_{t-1}$:

$$W_j(t) = \begin{cases} s_j^{(a)}(t) + \beta_j(X(t)) + \gamma_j(Y(t)), & \text{if } j \in \mathscr{A}_t \\ s_j^{(i)}(t;0) + \beta_j(X(t)) + \gamma_j(Y(t)), & \text{if } j \in \mathscr{I}_{t;0} \\ \beta_j(X(t)) + \gamma_j(Y(t)), & \text{if } j \in \mathscr{D}_t \end{cases} \tag{7.3.22}$$

For a dependent individual $j \in \mathscr{I}_{t-1}$ that moved into dependency at time $T^{(a)} < t$:

$$W_j(t) = \begin{cases} s_j^{(i)}(t; t-T^{(a)}) + \beta_j(X(t)), & \text{if } j \in \mathscr{I}_t \\ \beta_j(X(t)), & \text{if } j \in \mathscr{D}_t \end{cases} \tag{7.3.23}$$

Figure 7.8 illustrates one simulation run in the 3-state framework, comparing an active person (left) to an individual moving into dependency at time $T^{(a)} = 15$ and Figure 7.9 illustrates the average payoff together with a 95% confidence interval.

The product is shown to be actuarially fair in Theorem 7.3.

Theorem 7.3 (Actuarial fairness 3-state framework)**.**
The fairness conditions (7.3.7)*,* (7.3.8) *and* (7.3.20) *imply that the current account value is actuarially fair for a dependent individual if, at each time* $t = T^{(a)}, \ldots, \omega - x_j$*:*

$$c_j^{(i)}(t; t-T^{(a)}) = \mathrm{E}_t\left[\sum_{k=t+1}^{\omega - x_j} v(t,k) W_j(k) \,\middle|\, j \in \mathscr{I}_{t;t-T^{(a)}} \right]. \tag{7.3.24}$$

Similarly, it is actuarially fair for an active individual as, at each time $t = 0, 1, \ldots, \omega - x_j$*:*

$$c_j^{(a)}(t) = \mathrm{E}_t\left[\sum_{k=t+1}^{\omega - x_j} v(t,k) W_j(k) \,\middle|\, j \in \mathscr{A}_t \right]. \tag{7.3.25}$$

Proof: At time $\omega - x_j$, we have $q_{\omega;z}^{(i)} = 1$ and the cash flows only consist of mortality credits. Fairness condition (7.3.8) is supposed to hold implying $c_j^{(i)}(\omega - x_j; z) = 0, \forall z$. Assume that (7.3.24) holds for time t. For a dependent person $j \in \mathscr{I}_{t-1}$, with time spent

in dependency $z = t - T^{(a)}$, we have:

$$\begin{aligned}
&\mathrm{E}_{t-1}\left[\sum_{k=t}^{\omega-x_j} v(t-1,k)W_j(k)\,\middle|\, j \in \mathscr{I}_{t-1;z-1}\right] \\
&\quad = v(t-1,t)\Big(\mathrm{E}_{t-1}[W_j(t) + \mathbb{1}_{j\in\mathscr{I}_t}\cdot c_j^{(i)}(t;z)\,|\, j\in\mathscr{I}_{t-1;z-1}]\Big) \\
&\quad = v(t-1,t)\Big(\mathrm{E}_{t-1}[\mathbb{1}_{j\in\mathscr{I}_t}\cdot s_j^{(i)}(t;z) + \beta_j(X(t))|\, j\in\mathscr{I}_{t-1;z-1}] + p^{ii}_{x_j+t-1;z}\cdot c_j^{(i)}(t;z)\Big) \\
&\quad = v(t-1,t)\Big(p^{ii}_{x_j+t-1;z}\cdot s_j^{(i)}(t;z) + \mathrm{E}_{t-1}[\beta_j(X(t))] + p^{ii}_{x_j+t-1;z}\cdot c_j^{(i)}(t;z)\Big) \\
&\quad = v(t-1,t)\Big(p^{ii}_{x_j+t-1;z}\cdot s_j^{(i)}(t;z) + q^{(i)}_{x_j+t-1}\cdot v(t-1,t)^{-1}c_j^{(i)}(t-1;z-1) + p^{ii}_{x_j+t-1;z}\cdot c_j^{(i)}(t;z)\Big) \\
&\quad = c_j^{(i)}(t-1;z-1)\,.
\end{aligned}$$

This proves (7.3.24) for $t-1$. For an active person $j\in\mathscr{A}_{t-1}$, we also have that $c_j^{(a)}(\omega - x_j) = 0$. Using (7.3.7) and (7.3.8), backward iteration enables to obtain:

$$\begin{aligned}
&\mathrm{E}_{t-1}\left[\sum_{k=t}^{\omega-x_j} v(t-1,k)W_j(k)\,\middle|\, j \in \mathscr{A}_{t-1}\right] \\
&\quad = v(t-1,t)\mathrm{E}_{t-1}[W_j(t) + \mathbb{1}_{j\in\mathscr{A}_t}\cdot c_j^{(a)}(t) + \mathbb{1}_{j\in\mathscr{I}_t}\cdot c_j^{(i)}(t;0)|\, j\in\mathscr{A}_{t-1}] \\
&\quad = v(t-1,t)\Big(\mathrm{E}_{t-1}[\mathbb{1}_{j\in\mathscr{A}_t}\cdot s_{(j)}^{(a)}(t) + \mathbb{1}_{j\in\mathscr{I}_t}\cdot s_j^{(i)}(t;z) + \beta_j(X(t)) + \gamma_j(Y(t))|\, j\in\mathscr{A}_{t-1}] \\
&\qquad + p^{ai}_{x_j+t-1}\cdot c_j^{(i)}(t;0) + p^{aa}_{x_j+t-1}\cdot c_j^{(a)}(t)\Big) \\
&\quad = v(t-1,t)\Big(p^{aa}_{x_j+t-1}\cdot s_j^{(a)}(t) + p^{ai}_{x_j+t-1}\cdot s_j^{(i)}(t;0) + \mathrm{E}_{t-1}[\beta_j(X(t))] \\
&\qquad + \mathrm{E}_{t-1}[\gamma_j(Y(t))] + p^{ai}_{x_j+t-1}\cdot c_j^{(i)}(t;0) + p^{aa}_{x_j+t-1}\cdot c_j^{(a)}(t)\Big) \\
&\quad = v(t-1,t)\Big(p^{aa}_{x_j+t-1}\cdot s_j^{(a)}(t) + p^{ai}_{x_j+t-1}\cdot s_j^{(i)}(t;0) + q^{(a)}_{x_j+t-1}c_j^{(a)}(t-1)v(t-1,t)^{-1} \\
&\qquad + p^{ai}_{x_j+t-1}\Big((c_j^{(a)}(t) - c_j^{(i)}(t;0)) + (s_j^{(a)}(t) - s_j^{(i)}(t;0))\Big) \\
&\qquad + p^{ai}_{x_j+t-1}\cdot c_j^{(i)}(t;0) + p^{aa}_{x_j+t-1}\cdot c_j^{(a)}(t)\Big) = c_j^{(a)}(t-1)\,.
\end{aligned}$$

□

Note that at time $t = T^{(a)}$, we have that $s_j^{(a)}(t) - s_j^{(i)}(t;0) = (1-\alpha_j)b_j(t)$. As in the 2-state framework, the payoff is split into a fixed part, mortality and morbidity credits in a way that we obtain a desired average payoff. For an active individual, this average

payoff is $b_j(t)$, while for a dependent individual it is increased to $\alpha_j \cdot b_j(t)$, where $\alpha_j > 1$ is a predetermined constant. In the 3-state framework, we need to separately look at active and dependent individuals, as they have different time patterns of average mortality and morbidity credits. Mortality credits are shared within the whole group. However, dependent individuals receive a larger share of these credits due to their higher mortality risk. Theorem 7.4 shows how to determine the fixed part of the payoff and the account values for active and dependent individuals, respectively.

Theorem 7.4 (Choice of $s_j^{(a)}(t)$, $s_j^{(i)}(t;t-T^{(a)})$)**.**
Consider an annual time grid $t \in \mathbb{N}$. For a dependent individual ($j \in \mathscr{I}_t$), we follow Theorem 7.2 and use (7.3.17) *to obtain for each $T^{(a)} = 1,2,\ldots,\omega - x_j - 1$:*

$$s_j^{(i)}(t;t-T^{(a)}) = \alpha_j \cdot \hat{s}_j^{(i)}(t;t-T^{(a)}), \text{ for } t = T^{(a)}+1, T^{(a)}+2,\ldots \tag{7.3.26}$$

and the corresponding account value at time $t \geq T^{(a)}$:

$$c_j^{(i)}(t;t-T^{(a)}) = \sum_{u=t+1}^{\omega-x_j} v(t,u) s_j^{(i)}(u;u-T^{(a)})\,. \tag{7.3.27}$$

An active individual ($j \in \mathscr{A}_t$) receives the fixed payoff $s_j^{(a)}(t)$ determined via the backwards iteration:

$$\begin{aligned}
s_j^{(a)}(\omega-x_j) &= \frac{b_j(\omega-x_j)}{1+q_{\omega-1}^{(a)}}\,,\\
s_j^{(a)}(t) &= \frac{1}{1+q_{x_j+t-1}^{(a)}}\Bigg(b_j(t)\cdot p_{x_j+t-1}^{ai}(1-\alpha_j) - q_{x_j+t-1}^{(a)} \sum_{u=t+1}^{\omega-x_j} v(t,u) s_j^{(a)}(u)\\
&\qquad + b_j(t)\cdot\alpha_j + p_{x_j+t-1}^{ai} \sum_{u=t+1}^{\omega-x_j} v(t,u)\left(s_j^{(a)}(u) + s_j^{(i)}(u;u-t)\right)\Bigg), \qquad (7.3.28)\\
&\text{for } t = \omega-x_j-1, \omega-x_j-2,\ldots,1\,.
\end{aligned}$$

At time $T^{(a)}$, we have that:

$$s_j^{(i)}(T^{(a)};0) = s_j^{(a)}(T^{(a)}) + (\alpha_j - 1) b_j(T^{(a)})\,. \tag{7.3.29}$$

Proof: For an active person, we can compute the expected value of (7.3.22) to obtain:

$$\begin{aligned}
\mathrm{E}[W_j(t) \,|\, j \in \mathscr{A}_t] &= \mathrm{E}[s_j^{(a)}(t) + \beta_j(X(t)) + \gamma_j(Y(t)) \,|\, j \in \mathscr{A}_t]\\
&= s_j^{(a)}(t) + \mathrm{E}[X_j(t) \,|\, j \in \mathscr{A}_t] + \mathrm{E}[Y_j(t) \,|\, j \in \mathscr{A}_t]\\
&= s_j^{(a)}(t) + c_j^{(a)}(t-1) v(t-1,t)^{-1} q_{x_j+t-1}^{(a)}\\
&\qquad + p_{x_j+t-1}^{ai}\Big((c_j^{(a)}(t) - c_j^{(i)}(t;0)) + (1-\alpha_j) b_j(t)\Big)\,. \qquad (7.3.30)
\end{aligned}$$

We use (7.3.17) to obtain $s_j^{(i)}(t;t-T^{(a)}) = \alpha_j \cdot \tilde{s}_j^{(i)}(t;t-T^{(a)})$ for $t = T^{(a)}+1, T^{(a)}+2, \ldots$ and for all $T^{(a)}$. We have

$$c_j^{(i)}(t;0) = \sum_{u=t+1}^{\omega-x_j} v(t,u) s_j^{(i)}(u;u-t). \tag{7.3.31}$$

If survivors want on average an annual payoff of $b_j(t)$, we need to set

$$s_j^{(a)}(t) + c_j^{(a)}(t-1)v(t-1,t)^{-1} q_{x_j+t-1}^{(a)} + p_{x_j+t-1}^{ai} \Big((c_j^{(a)}(t) - c_j^{(i)}(t;0)) + (1-\alpha_j) b_j(t) \Big) = b_j(t).$$

We can iteratively solve this set of equations backwards in time to obtain (7.3.28). Equation (7.3.29) takes into account the immediate increase of benefits at the first payment date after moving into dependency. □

As in the 2-state case, the computation of the fixed payoff $s_j^{(a)}(t)$, $s_j^{(i)}(t;t-T^{(a)})$ in Theorem 7.4 can be carried out for each individual separately.

7.4 Conclusion

A life-care tontine relies on the natural hedge inherent between mortality and morbidity risks. When moving into dependency, individuals may need a higher payoff for a shorter remaining lifetime, allowing to easily pool these risks with healthy individuals.

As in the case of a life-care annuity, the pooling of mortality and morbidity risks reduces adverse selection costs and provides more people access to long-term care insurance. The advantage of a tontine scheme is an additional reduction in adverse selection costs driven by the uncertainty of future tontine cash-flows (see, e.g., Valdez et al. (2006)). Further, the insurance provider is merely administrative, leading to a significant reduction in risk and administration charges (see, e.g., Chen, Hieber & Klein (2019)). The drawback naturally is that the systematic risk lies with the policyholders.

A major innovation is the development of a creative product design: cashflows can be smoothed to fit the current and future needs of the market. The product is actuarially fair *at each point in time*, allowing people to later join the tontine scheme. The individual flexibility of our payoff design answers the individual practical needs of the insureds.

Technically, we rely on a backward iteration used to deduce the smoothed cashflows patterns and the separation of cash-flows in a fixed withdrawal, mortality and morbidity credits. The flexibility and fairness of the system results from the fact that this iteration can be carried out individually for each pool member. The pooling scheme shares the mortality and morbidity *risks* within the pool. The average future payoffs and shares in mortality and morbidity credits are, however, based on each individual's risk, for example the age and health status. The 2-state and 3-state models are applied to real life data, providing coherent simulations and illustrating the adequacy of our product framework.

There exist alternatives to fully funded solutions for LTC insurances. Solutions could also comprise a risk transfer to the financial market or to the active population via social security and are discussed in the 'Discussions and Extensions' Chapter.

References

Bernhardt, T. & Donnelly, C. (2019). Modern tontine with bequest: innovation in pooled annuity products. Insurance: Mathematics and Economics 86, 168-188.

Brown, J. & Warshawsky, M. (2013). The Life Care Annuity: A New Empirical Examination of an Insurance Innovation that Addresses Problems in the Markets for Life Annuities and Long-Term Care Insurance. The Journal of Risk and Insurance 80(3), 677-704.

Chen, Y. (2003) Applications of the trade-off principle in both public and private sectors. Journal of aging and health 15(1), 15-44

Chen, A., Hieber, P. & Klein, J. (2019). Tonuity: A novel individual-oriented retirement plan. ASTIN Bulletin 49(1), 5-30.

Chen, A., Hieber, P. & Rach, M. (2020). Optimal retirement products under subjective mortality beliefs. Insurance: Mathematics and Economics.

Chen, A., Hieber, P. & Nguyen, T. (2019). Constrained non-concave utility maximization: An application to life insurance contracts with guarantees. European Journal of Operational Research 273(3), 1119-1135.

Chen, A., Rach, M. & Sehner, T. (2020). On the optimal combination of Annuities and Tontines. ASTIN Bulletin 50(1), 95-129.

Chen, A., Fuino, F., Sehner, T. & Wagner, J. (2020) Valuation of Long-Term Care Options Embedded in Life Annuities. working paper.

Denuit, M. & Dhaene, J. (2012). Convex order and comonotonic conditional mean risk-sharing. Insurance: Mathematics and Economics 51, 265-270.

Denuit, M., Lucas, N. & Pitacco, E.(2019) Pricing and Reserving in LTC Insurance Actuarial Aspects of Long Term Care. Springer, 129-158.

Denuit, M. (2019). Size-biased transform and conditional mean risk sharing, with application to P2P insurance and tontines. ASTIN Bulletin 49(3), 591-617.

Denuit, M. & Robert, C. (2020). From risk sharing to pure premium for a large number of heterogeneous losses (No. 2020015). UC Louvain, Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA).

Donnelly, C., Guillén, M. & Nielsen, J. (2013). Exchanging uncertain mortality for a cost. Insurance: Mathematics and Economics 52(1), 65-76.

Donnelly, C., Guillén, M. & Nielsen, J. (2014). Bringing cost transparency to the life annuity market. Insurance: Mathematics and Economics 56, 14-27.

Forman, J. & Sabin, M. (2015). Tontine pensions. University Pennsylvania. Law Review 163, 755-831.

Forman, J. & Sabin, M. (2016) Survivor funds. Pace Law Review 37, 204-291.

Fullmer, R. & Sabin, M. (2018) Individual tontine accounts. Available at SSRN 3217551.

Fries, J. (1980). Aging, natural death, and the compression of morbidity. The New England Journal of Medicine 303(3), 245-250.

Fuino, M. & Wagner, J. (2020) Duration of long-term care: socio-economic factors, type of care interactions and evolution. Insurance: Mathematics and Economics 90, 151-168.

Friedman, B. & Warshawsky, M. (1990). The cost of annuities: Implications for saving behavior and bequests. The Quarterly Journal of Economics 105(1), 135-154.

Hansen, M. & Miltersen, K. (2002) Minimum rate of return guarantees: the Danish case. Scandinavian Actuarial Journal (4), 280-318.

Hieber, P., Natolski, J. & Werner, R. (2019). Fair valuation of cliquet-style return guarantees in (homogeneous and) heterogeneous life insurance portfolios. Scandinavian Actuarial Journal (6), 478-507.

Holzmann, R., Alonso-Garcia, J., Labit-Hardy, H. & Villegas, A. (2019). NDC schemes and heterogeneity in longevity: Proposals for redesign. In R. Holzmann, E. Palmer, R. Palacios, and S. Sacchi (Eds.), Progress and Challenges of Nonfinancial Defined Pension Schemes - Volume 1: Addressing Marginalization, Polarization, and the Labor Market, Chapter 16. Washington, D.C.: The World Bank.

Li, Y. & Rothschild, C. (2019) Selection and redistribution in the Irish tontines of 1773, 1775, and 1777. Journal of Risk and Insurance 87(3).

Milevsky, M. & Salisbury, T. (2015). Optimal retirement income tontines. Insurance: Mathematics and Economics 64, 91-105.

Milevsky, M. & Salisbury, T. (2016). Equitable retirement income tontines: Mixing cohorts without discriminating. ASTIN Bulletin 46(3), 571-604.

Murtaugh, C., Spillman, B. & Warshawsky, M. (2001). In sickness and in health: An annuity approach to financing long-term care and retirement income. Journal of Risk and Insurance, 225-253.

OECD (2020) Long-term Care and Health Care Insurance in OECD and Other Countries Available at www.oecd.org/fin/insurance/Long-Term-Care-Health-Care-Insurance-in-OECD-and-Other-Countries.htm

Piggott, J., Valdez, E. A. & Detzel, B. (2005). The simple analytics of a pooled annuity fund. Journal of Risk and Insurance 72(3), 497-520.

Pitacco, E. (2002) LTC insurance in Italy. In: 2002 Health Seminar on Critical Issues in Managing Long-Term Care Insurance. XXVII ICA - Cancun.

Qiao, C. & Sherris, M. (2013) Managing systematic mortality risk with group self-pooling and annuitization schemes. Journal of Risk and Insurance 80 (4), 949-974.

Rickayzen, B. (2007). An Analysis of Disability-linked Annuities, Actuarial Research Paper No. 180 Faculty of Actuarial Science and Insurance Cass Business School. City University, London.

Sabin, M. (2010) Fair tontine annuity. Available at SSRN 1579932.

Schumacher, J. (2018). Linear versus nonlinear allocation rules in risk sharing under financial fairness. ASTIN Bulletin 48, 995-1024.

Shao, A., Sherris, M. & Fong, J. (2017). Product pricing and solvency capital requirements for long-term care insurance. Scandinavian Actuarial Journal 2, 175-208.

Shi, P. & Zhang, W. (2013). Managed care and health care utilization: Specification of bivariate models using copulas. North American Actuarial Journal 17(4), 306-324.

Spillman, B., Murtaugh, C. & Warshawsky, M. (2003). Policy Implications of an Annuity Approach to Integrating Long-Term Care Financing and Retirement Income. Journal of Aging and Health 15(1), 45-73.

Stamos, M. (2008) Optimal consumption and portfolio choice for pooled annuity funds. Insurance: Mathematics and Economics 43 (1), 56-68.

Valdez, E., Piggott, J. & Wang, L. (2006). Demand and adverse selection in a pooled annuity fund. Insurance: Mathematics and Economics 39 (2), 251-266.

Vidal-Melia, C., Ventura-Marco, M. & Pla-Porcel, J. (2020) An NDC approach to helping pensioners cope with the cost of long-term care. Journal of Pension Economics & Finance 19(1), 80-108.

Weinert, J.-H. & Gründl, H. (2017) The modern tontine: An innovative instrument for longevity risk management in an aging society. Available at SSRN 3088527.

Part IV

Discussion and Extensions

Chapter 8

Discussion and extensions

Population ageing has produced a higher number of individuals exposed to the risk of becoming sick or dependent. There is an increasing demand for the long-term care needs of individuals. The issue of forecasting morbidity in the long-term is delicate. Several different systematic risks are combined. For example the morbidity risk is the risk arising from uncertainty in future morbidity trends. The longevity risk or the progressive increase in lifetime duration is a demographic aspect that can both impact disabled lives and healthy lives. Inflation is a particular systematic risk and is the focus of Chapter 2.

An adequate management of longevity, morbidity and inflation risks requires insurance companies and pension plans to model and measure them. Until recent years, both mortality and morbidity have been traditionally modeled in a deterministic framework, while the insurance companies' ability to read demographic trends has significantly improved over the last decade.

The aim of this thesis is to analyse and disentangle health insurance related risks and their interactions and also propose risk management strategies. In the first part stochastic models have been introduced to better capture the health expenditures evolution and their frequency and severity components in a Lee-Carter or time-to-death approach. Our work was motivated by the similar age structure generally observed for health insurance claim frequencies and yearly aggregate losses on the one hand and mortality on the other hand, as stated in Chapter 3. Proximity-to-death is confirmed in Chapter 4 to be a prominent cost factor compared to attained age. We refer the interested reader to Fong (2017) for the distinction between a 'chronological age' and a latent 'physiological age', explaining the natural heterogeneity in health risks. We introduced longevity dynamics in the models, allowing to link morbidity to mortality projections. We showed that longevity and medical inflation effects can be disentangled with the help of an appropriate frequency-severity decomposition.

The KBC product in Chapter 6 constitutes an innovative risk hedging design. The KBC package consists of a whole-life insurance coverage together with an LTC annuity

with monthly payment m, subject to a deferred period. The interest of combining longevity and morbidity risks is clear. The length of the deferred period depends on policyholder's age x at policy issue. The product is sold as a combination of a life insurance contract and a LTC cover. Depending on the country, these two products may fall under different lines of business and must then be managed separately.

Two approaches have been proposed when facing systematic health insurance related risks: the ex-post premium adaptation (cf. Chapter 2) or the collaborative 'risk pooling' mechanism (cf. Chapter 7). As it was said, an insurer is not entitled to bear on its own the systematic risk that is associated with lifelong health insurance private contracts. The low interest rate environment, the demography and the introduction of Solvency II has strengthened the trend to shift more and more risks from the insurance provider to the consumer or to an external party. This actually not only affects the private insurance market but also occupational and state pensions, moving from a defined benefit-type guarantee to collective defined contribution.

In order to share systematic mortality and morbidity risks through an affordable insurance scheme, solutions including mutual insurance, financial securitization or PAYG system, could be considered.

Although materializing only slowly, some interesting developments have occurred around mortality-linked securities, and these could be transposed to morbidity-linked and LTC-linked products. But the securitisation of the LTC risk is at present non-existent.

Another solution is the risk transfer to the active population. For example a Non-financial Defined Contribution (NDC) pension system could include long-term care in its mechanism, see Vidal et al. (2020). The interesting idea of combining retirement and long-term care contingencies is not new. For example we could relate to Chen (2003), who advocates the creation of a 'social security long term care' or Pitacco (2002) who proposes enhanced pension annuities (EPA) funded by contributions.

As nicely resumed in Vidal et al. (2020), 'the solution to long-term care finance needs to be found by having an integrated vision of the roles of the Market, State and family'.

For its part Chapter 7 proposes a mutual-risk-sharing scheme known as tontines. Tontines at present are not particularly frequent but they do exist, e.g. TIAA-CREF retirement fund in the USA. In Belgium *les clauses de tontine ou d'accroissement en pleine propriété* do exist (see notaire.be). In Australia the group Mercer proposes the Lifetimeplus solution, which has several tontine aspects.

In France 'Le Conservateur' is a mutualist group founded in 1844 and is a pioneer in the tontine market. They advertise very nice returns but warn that this investment should be considered as a diversification tool in addition to a life insurance policy or another investment plan. For example, the last tontine of the current Conservator over the period 1997-2017 generated an average annual return of 4.52% (net of contract management fees, excluding tax and social security contributions), in the case of a subscription at the age of 45 over 15 years in the form of a one-off payment (cf. conservateur.fr).

In 2014, Forman & Sabin (2014) argued that tontine pensions would have two major advantages over traditional pensions, as they would always be fully funded, and the plan

sponsor would not be required to bear the investment and actuarial risks, see also Milevsky (2015).

In 2017, the brand Tontine Trust is delivering secure low cost Tontine Pensions based on the Forman and Milevsky framework. This fintech company is based in Ireland and is providing pension benefits on a digital platform. In 2019, Fullmer and Sabin introduced the concept of individual tontine accounts (ITAs). They also showed that it is possible to engineer payouts within a tontine structure that are immune to interest rate and reinvestment risk.

The life-care tontine design proposed in Chapter 7 has highlighted the interest of pooling mortality and morbidity risks. Yet several simplifying assumptions might be relaxed in future research. We notably assumed that each individual's mortality and morbidity risks are independent of the other pool members' risks. This is only true if there are no systematic risks affecting every pool member simultaneously, like a pandemic, improved medication or a general increase in life expectancy. The existence of systematic morbidity risk is still controversially discussed (for example Fries (1980) detects a rectangularia-tion of morbidity while Fuino (2020) finds that "the duration of LTC has not significantly changed in the period from 1995 to 2009").

The topic of transferability of reserves in health insurance has briefly been approached in the reserve computations of chapter 2 and 6. The matter is delicate and we refer the interested reader to Dhaene (2015). Non-transferable reserves bind the insured to the insurer, reduce competition and imply inevitably some adverse selection. Actually only a credible reserve, i.e. one that reflects the state of health of the insured party who wishes to cancel his policy as deducted from the costs of past claims, can be used as a basis for to the calculation of a redemption value. Failing to do so, healthy policyholders would be favoured by an higher transfer value and their removal from the portfolio would break the mutualisation, putting the solvency of the insurer at risk. The surrender value should therefore be lower for the insured persons with a favourable (probably healthy) claim history, and higher for others.

Different papers have tried to adapt credibility theory in the health insurance domain (see Lu et al. 2012, Fong et al. 2015). The idea of using Hidden Markov Models, also called frailty model in survival analysis, is of particular interest, although the parameter estimation of such models is complex as the observations are often insufficient with regards to the number of parameters and can include zero values.

References

- Dhaene, J., Godecharle, E., Antonio, K. & Denuit, M. (2015) On the transferability of reserves in lifelong health insurance contracts. ASTIN, AFIR/ERM and IACA Colloquia, Sydney.

- Holzmann, R., Alonso-Garcia, J., Labit-Hardy, H. & Villegas, A. (2019). NDC schemes and heterogeneity in longevity: Proposals for redesign. In R. Holzmann, E.

Palmer, R. Palacios, and S. Sacchi (Eds.), Progress and Challenges of Nonfinancial Defined Pension Schemes - Volume 1: Addressing Marginalization, Polarization, and the Labor Market, Chapter 16. Washington, D.C.: The World Bank.

- Forman, J. & Sabin, M. (2015). Tontine pensions. University Pennsylvania. Law Review 163, 755-831.

- Fong, J. H., Sherris, M. & Yap, J. (2017). Forecasting disability: Application of a frailty model. Scandinavian Actuarial Journal, 2017(2), 125-147.

- Fries, J. F. (1980). Aging, natural death, and the compression of morbidity. The New England Journal of Medicine, 303(3), 245-250.

- Milevsky, M. & Salisbury, T. (2015). Optimal retirement income tontines. Insurance: Mathematics and Economics 64, 91-105.

- Fuino, M. & Wagner, J. (2020) Duration of long-term care: socio-economic factors, type of care interactions and evolution. Insurance: Mathematics and Economics 90, 151-168.

- Lu, Y. & Zeng, L. (2012) A nonhomogeneous Poisson hidden Markov model for claim counts. Astin Bulletin 42(01), 181-202.